# В.З. Аладьев, В.Н. Харитонов

## Курс Общей Теории Статистики

❖

*Fultus™ Books*

# General Theory of Statistics

by

V. Z. Aladjev and V.N. Haritonov

ISBN 1-59682-086-1



Published by Fultus Publishing

Publisher Web Site: *www.fultus.com*

Fultus eLibrary: *elibrary.fultus.com*

Online Book Superstore: *store.fultus.com*

Writer web site: *writers.fultus.com/aladjev/*

# Оглавление

# От Авторов

Предмет ***общей теории статистики*** входит в число вузовских дисциплин, определяющих общепрофессиональную подготовку специалистов по целому ряду специальностей из области так называемых *общественных* или *социальных* наук. Статистика, как общественная наука, имеет целью дать студентам вузов представление о содержании экономического или социального явления, познакомить с ее основными понятиями, методологией и методиками расчета важнейших статистических аналитических показателей. ***Общая теория статистики*** разрабатывает приемы количественного анализа, методы сбора, исследования экономико-социальной информации, позволяет оценить динамику и последовательность исследуемых объектов, процессов и явлений, и т.д. В число основных задач данного предмета входят такие как: изучение общих свойств массовых явлений и методов их анализа, раскрытие содержания и конкретных методов построения системы показателей для характеристики образа жизни населения и различных аспектов социально- экономических отношений. В концептуальном плане предмет статистики связан с философией в теоретико-методологическом отношении, и с математикой (*теория вероятностей и математическая статистика*) – в методическом. Изучение данной дисциплины создает хорошие предпосылки для качественного освоения экономического анализа, основ экономической теории, бухгалтерского учета и целого ряда других экономико-социальных дисциплин.

Предметом статистики является *количественная* сторона массовых экономико-социальных явлений в их неразрывной связи с *качественной* стороной конкретных условий места и времени. Это определяет и основные черты предмета ***общей теории статистики,*** а именно: (**1**) *социально-общественный* характер статистики, (**2**) акцент на *количественную* сторону социально-общественных явлений (*в отличие от других общественных наук*), (**3**) ориентация на исследование массовых явлений, (**4**) исследование *количественной* стороны явлений совместно с их *качественной* стороной на основе системы статистических показателей и (**5**) исследование количественной стороны явлений в конкретных условиях места и времени. Предлагаемая книга и рассматривает основные вопросы предмета общей теории статистики.

Настоящая книга представляет собой статистический курс для начинающих студентов во всех областях социально-экономических наук. Книга представляет собой пособие по курсу ***"Общая Теория Статистики ",*** включая ряд не совсем традиционных тем. Прежде всего, это касается математических основ статистики и использования компьютерных технологий в статистическом исследовании. Тематический выбор глав и разделов книги вызван не только интересами и вкусами авторов, но также и современными тенденциями в прикладной статистике и ориентации данной работы на студентов социальных и экономических наук.

С той или иной степенью детализации рассматриваются основные темы общей теории статистики, а именно: (**1**) *предмет и метод статистики*, *ее место среди других наук*, (**2**) *элементы теории вероятностей*, (**3**) *основания статистического наблюдения*, (**4**) *группировка*, *сводка*, *и представление статистических данных*, (**5**) *абсолютные и относительные величины*, (**6**) *основы метода средних*, (**7**) *корреляционный и регрессионный анализ*, (**8**) *элементы анализа вариационных и временных рядов*, (**9**) *элементы индексного метода*, (**10**) *компьютерные средства статистического анализа данных*.

Книга содержит ряд конкретных предложений об усовершенствовании статистической практики; многие из этих предложений базируются на нашем опыте практической работы в статистических органах СССР и Эстонии в начале девяностых годов прошлого столетия, а также в банковской системе Эстонии. В этом отношении может быть достаточно интересен аспект сравнения советской статистической школы, унаследовавшей весьма много традиций всемирно известных российской и советской школ теории вероятностей и статистики, с западной статистической наукой.

Данная книга написана для достаточно широкой аудитории преподавателей, студентов университетов и колледжей, исследователей, статистиков, а также пользователей статистики в поведенческих и социальных науках. Прежде всего, книга ориентирована на широкий круг читателей, изучающих статистические дисциплины в университетах и колледжах; однако, она может быть полезна также для читателей, самостоятельно изучающих статистику. Присутствие в книге ряда нетрадиционных тем делает ее полезным руководством для всех тех, кто в своей профессиональной деятельности имеет дело со статистическим анализом данных различных природы и характера.

Новые особенности этого курса – многочисленные примеры статистических процедур, реализованных в среде известного пакета *Maple*. Исходные тексты некоторых из процедур включены в книгу, что позволяет непосредственно использовать их в среде пакета *Maple* с конкретными данными читателя. Эти и другие процедуры находятся в пользовательской Библиотеке, позволяя выполнять простой статистический анализ данных различного характера в среде пакета *Maple*. Данная Библиотека может быть бесплатно загружена со следующих вебсайтов:

**http://www.aladjev.newmail.ru/DownLoad/UserLib6789.zip**
**http://writers.fultus.com/aladjev/source/UserLib6789.zip**

Материал настоящей книги базируется на трех наших российских книгах, тираж которых был полностью распродан, и английской книге [347]. Эти книги были написаны на основе ряда курсов лекций по Общей Теории Статистики, Теории Вероятностей и Математической Статистики для студентов Университетов Белоруссии и Балтики, которые специализируются в области экономических и социальных наук (*экономика*, *международное право*, *юриспруденция*, *политология*, *социология*, *психология*, *банковское дело*, *бухгалтерия и т.д.*). Книга содержит весьма обширную как русскую, так и английскую литературу по различным аспектам статистики. Наряду с этим, книга снабжена *полезным* списком статистических *организаций*, статистических периодических изданий и т.д.

Принимая во внимание все более усиливающуюся роль статистического анализа (*как одной из главных предпосылок обеспечения обратной связи в управлении*) для обеспечения достоверной информации в *экономике*, *финансах*, *банковском деле*, *бизнесе* и *управлении*, есть все основания полагать, что предлагаемая книга найдет многочисленных читателей различного уровня и сфер деятельности.

# Предисловие

*Статистическая грамотность* составляет неотъемлемую составную часть профессиональной подготовки каждого экономиста, финансиста, социолога, политолога, клерка, а также любого специалиста, имеющего дело с различным анализом массовых явлений, будь то социально-общественные, экономические, технические, научные и др. Работа этих групп специалистов неизбежно связана со сбором, разработкой и анализом данных *статистического* (*массового*) характера. Нередко им самим приходится проводить статистический анализ различных типа и направленности либо знакомиться с результатами статистического анализа, выполненного другими. В настоящее время от специалиста, занятого в любой области науки, техники, производства, бизнеса и др., связанной с изучением массовых явлений, требуется, чтобы он был по крайней мере статистически грамотным человеком. В конечном счете невозможно успешно специализироваться по многим дисциплинам без освоения какого-либо уровня статистического курса. Поэтому большое значение имеет знакомство с общими категориями, принципами и методологией статистического анализа данных.

Основу статистической грамотности в значительной мере дает предмет ***"Общая теория статистики"***, содержание которого во многих учебных программах вузов не отвечает современному уровню развития статистической методологии и ее прикладным аспектам. Поэтому настоящая книга является попыткой представить наш взгляд на содержание курса общей теории статистики, как ядра всей статистической методологии. Предполагается, что статистические *знания*, приобретенные в рамках данного курса, послужат хорошей *отправной* точкой для последующего изучения курсов *специальных* статистик: *экономической*, *отраслевых*, *медико-биологической*, *судебной*, *демографической*, *банковской* и др.

Работу над книгой затрудняло, помимо сложности и обширности предмета, отсутствия по ряду вопросов общепринятых положений (*включая само определение термина* ***"статистика"***), то обстоятельство, что имеются существенные различия в учебных планах вузов финансово-экономических, социальных и общественных дисциплин, а также различные взгляды на статистику, как предмет преподавания. В свете современных требований и тесной связи общей статистики с ее математической основой – *теорией вероятностей* и *математической статистикой* – представляется целесообразным включить в курс элементы этих наук. Одной из причин быстрого роста статистических исследований в последние десятилетия является все возрастающая легкость обработки больших числовых массивов средствами современной *вычислительной техники* (**ВТ**), особенно с появлением класса *персональных компьютеров* (**ПК**), что позволило *непосредственно* производить статистический *анализ* (*по крайней мере первичный*) на месте сбора статистических данных. Данное обстоятельство сделало целесообразным представить обзор программных средств, ориентированных на статистический анализ данных, и дать представление о их возможностях и принципах использования.

При написании данной книги мы постоянно имели в виду цель достижения всеобщей статистической грамотности, поэтому наша задача состояла в представлении основных концепций статистики как методологической вспомогательной науки (*методологии*), целью которой является разработка методов сбора, обработки и анализа числовых данных, их интерпретация и проникновение в структуру и суть массовых явлений, в первую очередь, социально-экономического характера.

Книга состоит из **10** глав и освещает обширный материал по общей теории статистики – от исторического экскурса, элементарных статистик до элементов теории вероятностей, регрессионного и корреляционного анализов, анализа вариационных и динамических рядов, элементов индексного метода и обсуждения программных средств статистического анализа. Однако ограниченный объем книги не позволил изложить рассмотренные вопросы курса с одинаковой полнотой. Отсюда – *строгое* освещение лишь сути дела без *обстоятельного* проведения, например, математических доказательств и рассмотрения смежных (*часто интересных и важных*) проблем. Хотя в ряде случаев и приводятся оригинальные решения. Основные положения материала сопровождаются соответствующими упражнениями и примерами, проработку которых мы считаем необходимой при добросовестной подготовке читателя к экзаменам.

Кратко о содержании отдельных глав книги. *Первая* глава рассматривает предмет и метод статистики, ее структуру и местоположение в ней общей теории статистики – ядра всех прикладных статистик и основы статистической грамотности. Даются краткий экскурс по истории статистики и основные принципы современной организации государственной статистики. Завершает главу обсуждение основных задач статистики, особенностей ее методологии, а также основные понятия и категории статистики.

Во *второй* главе рассматриваются элементы теории вероятностей – теоретической основы математической статистики и базирующихся на ней прикладных статистиках. Вводится классическое определение вероятности, определяются случайные величины и основные законы их распределения. Представлены наиболее важные критериальные распределения, используемые в статистике, и ряд интересных примеров на закрепление изучаемого материала. Представленные сведения позволяют не только решать целый ряд полезных вероятностных задач, но и составляют первичную основу вероятностного метода, а также используются в дальнейшем изложении.

*Третья* глава представляет *основы* статистического *наблюдения* – *первого* этапа статистического анализа: программа и план наблюдения; основные формы, виды и способы наблюдения, а также вопросы точности наблюдения и специальные вопросы отчетности и переписи, имеющие большое практическое значение.

В *четвертой* главе представлены основы второго этапа статистического анализа – сводка, группировка и представление статистических данных. Рассматриваются задачи сводки, различные типы *группировок* и *классификаций*, а также статистические ряды распределения, табличное и графическое представление статистических данных.

В *пятой* главе рассматриваются абсолютные и относительные статистические величины. Тогда как *шестая* глава представляет элементы *метода средних величин* – одного из важнейших приемов обобщения статистических данных. Детально обсуждается базовое понятие *средней арифметической* и ряда других типов *средних*, включая структурные средние; рассматривается методика выбора типа средней для анализа тех или иных статистических совокупностей.

*Седьмая* глава представляет элементы анализа вариационных рядов. В ней рассматриваются такие вопросы как: *показатели* и *меры вариации* признаков; анализ *формы кривой распределения*; элементы теории выборочного метода, а также корреляционного и регрессионного анализа зависимостей между исследуемыми явлениями. При этом, представлены некоторые новые взгляды и соображения на методологию корреляционного анализа, как метода выявления связей (*зависимостей*) между исследуемыми явлениями.

В *восьмой* главе рассматриваются элементы анализа временных рядов. Изложение начинается с типов динамических (*временных*) рядов, их построения и представления, а также *определения*

основных статистических показателей рядов. Затем рассматриваются средние показатели и вопросы выявления основной тенденции (*тренда*) ряда. Изложение завершается анализом случайной компоненты и исследованием периодических колебаний ряда. В качестве некоего иллюстративного материала исследуются динамические ряды, отражающие цитируемость публикаций (*Таллиннской Творческой Группы*) **ТТГ** по математической теории однородных структур (*Cellular Automata*) и ее приложениям в СССР, за рубежом и в целом. Полученные в ходе *анализа* результаты позволяют делать весьма интересные *выводы* о *творческой* активности группы.

Индексному методу анализа посвящена *девятая* глава книги. Обсуждается понятие индексов, их типы и назначение. Рассматриваются индивидуальные и агрегатные индексы, а также *средние*, *цепные* и *базисные* индексы. Глава завершается обсуждением важнейших *экономических* индексов и их взаимосвязей. Вводится ряд показателей и индексов для анализа научной активности, которые могут оказаться весьма полезными в ряде разделов современного науковедения. Из-за обширности данного раздела современной статистики, основная доля которого приходится на отраслевые статистики, в главе представлены только основные элементы индексного метода анализа.

Наконец, в *десятой* главе обсуждаются компьютерные средства статистического анализа данных. Дается краткий обзор такого типа программного обеспечения, рассматривается использование класса *персональных компьютеров* (**ПК**) для решения статистических задач, а также иллюстрируются элементы статистического анализа данных в среде весьма известного математического пакета *Maple*. Подобный материал, на наш взгляд, должен отныне являться неотъемлемой составной частью курса общей теории статистики, ибо основным орудием труда современного статистика любого уровня (*либо специалиста, по роду своей деятельности имеющего дело со статистическим анализом различного рода данных*) все шире становится персональный компьютер либо другой тип *вычислительной техники* (**ВТ**).

Большинство примеров, иллюстрирующих те или иные положения, методы и приемы статистики, базируются на первичных статистических данных – совокупности научно-прикладных публикаций **ТТГ** за период 1970-2000 г.г. Общепонятная суть такой *совокупности* позволяет не отвлекаться на специфические понятия и сконцентрировать свое внимание на сугубо статистических вопросах. Наряду с этим используемая методика анализа такого типа совокупностей может оказаться полезной при статистическом анализе научной активности и стимулировать исследования в данном направлении, составляющем самостоятельную ветвь современного науковедения в целом.

Приведенная в книге обширная отечественная и зарубежная библиография охватывает основные разделы общей теории статистики и ряд связанных с ними направлений, что позволяет читателю выбирать наиболее подходящую для его целей литературу обзорного, монографического, методического, методологического, справочного или учебного *характера*.

Настоящая книга представляет расширенное изложение материала курса лекций по общей теории статистики, которые были прочитаны первым автором в июне 1995 в Гродненском филиале Минского Института Современных Знаний (*Западная Белоруссия*). Основная цель данного курса состояла в том, чтобы представить основные принципы современной теории статистики, ориентируемой, прежде всего, на студентов в области социальных, финансовых, экономических и не естественных наук, в целом. Авторы предлагают читателям данную книгу и будут весьма благодарны и признательны за их предложения и критические замечания, высланные по одному из адресов, указанных на вебсайте:

**http://www.aladjev.newmail.ru**

# Глава 1.
# Предмет и метод статистической науки

## 1.1. Предмет статистики и его местоположение

Изучение *истории статистики* является важнейшей составляющей экономико-финансового и статистического образования, включающая описание процесса возникновения и развития статистического учета, анализа и теории статистики, ее концепции. Развитие статистики, *в первую очередь*, определяется потребностями развития общества и государства, их социально-экономическими потребностями. Даже краткие сведения по истории статистики позволяют лучше понять *суть* и *цели* данного научно-прикладного предмета, имеющего самые широкие приложения. Цель данного раздела состоит как в историческом обзоре развития предмета, так и в ознакомлении с основными его аспектами, тем более, что большинство из них не рассматриваются в настоящей книге.

Статистические обработка и анализ данных восходят к глубокой древности человечества, тогда как сам термин ***"статистика",*** по-видимому, происходит более или менее косвенно от латинского слова ***"Status",*** имеющего в средневековой латыни значение "политическое состояние". Появившись в науке, этот термин проходит сложный путь эволюции, не имея до сих пор общепринятого определения. Это не удивительно, ибо статистика формируется на стыке ряда наук и прикладных методологий учета, контроля и анализа данных, и имеет весьма обширные приложения. В работе [88] приводится около 200 определений термина ***"статистика",*** принадлежащих крупным ученым (*начиная с* 17 *века и до настоящих дней*) и позволяющих более многогранно взглянуть на сам предмет статистики. Даже в современных толковых словарях [107] статистика определяется весьма многозначно: **(1)** наука, изучающая количественные показатели развития как общества, так и общественного производства; **(2)** количественный учет всякого рода массовых явлений; **(3)** научный метод количественных исследований в некоторых областях знания. На наш взгляд, можно выделить ***три*** основных этапа в эволюции смыслового значения ***"статистика".*** Сначала оно означало учение об экономическом и политическом состоянии государства, базирующемся на числовом анализе его экономических характеристик. На втором этапе под *"статистикой"* стали понимать обработку числовых данных *любой* природы. Сегодня под термином *"статистика"* понимаем приложение методов математической статистики в различных сферах человеческой деятельности и, в первую очередь, в социально-общественных и экономических науках.

Многие современные авторы и, в первую очередь, советские [64-69, 71, 72] косвенно или явно относят статистику к общественным наукам. По нашему мнению, статистика вовсе не является наукой в строгом понимании этого термина. Нам представляется следующим взаимоотношение статистики и других математических наук (рис. **1**), которое носит вполне субъективный характер, но которое сформировалось не только в процессе академической деятельности с использованием результатов и методов статистического анализа, но и в связи с длительной работой в области прикладной статистики в системе Главного статистического управления ЦСУ ЭССР, Эстонской республики и ВГПТИ ЦСУ СССР.

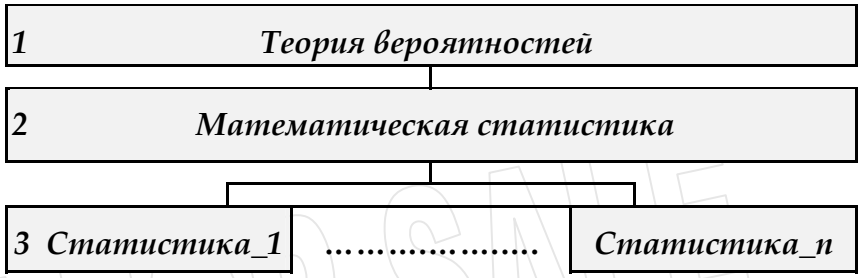



**Рис. 1. Место предмета статистики в системе других наук**

Строгой математической основой для исследования вероятностных закономерностей в абстрактном виде является ***теория вероятностей*** (на рис. **1;** *Уровень* **1**). Подобно тому, как математический анализ (*и особенно теория интегро-дифференциальных уравнений*) является основным математическим аппаратом при изучении физических закономерностей, теория вероятностей представляет наиболее пригодный инструмент для исследования процессов и явлений, подверженных действию *случайных факторов*. Как и все математические результаты, положения теории вероятностей носят абстрактный, безразличный к конкретной природе массовых явлений характер.

Теория вероятностей служит фундаментом ***математической статистики*** – *самостоятельной* прикладной области современной математики (на рис. **1;** *Уровень* **2**). Специфическая задача математической статистики – получать сводные абстратно-количественные характеристики *совокупности* данных при абстрагировании от реального содержания явления, описываемого этими данными. Поэтому при определенных условиях категории и методы математической статистики применимы к исследованию *массовых явлений* в различных областях естественных и общественных наук.

Собственно ***статистика*** (на рис. **1;** *Уровень* **3**) представляет собой некоторую совокупность самостоятельных статистик, содержащих методики и методологии сбора, анализа данных и интерпретации результатов в конкретной прикладной области (*физические и химические эксперименты*, *медико-биологические исследования*, *экономика*, *финансы*, *социология и др.*). Таким образом, в качестве предметов ***"Статистика_1"***, ***... ,"Статистика_n"*** можно представить себе такие дисциплины, как биометрия, эпидемиологическая статистика, экономическая статистика, банковская статистика и т.д. При этом, каждая компонента **3**-го уровня иерархии (рис. **1**) допускает отдельную детализацию, уточняющую объект статистического анализа. Таким образом, целью конкретной статистики является: (**1**) разработка *обоснованных* методик сбора числовых совокупностей, характеризующих конкретное явление или отдельную его составляющую; (**2**) разработка классификаторов, четких понятий и определений; (**3**) выбор метода анализа совокупности из числа имеющихся в математической статистике или формулировка новых математических проблем и (**4**) интерпретация результатов статистического анализа, а при необходимости и разработка прогнозов развития исследуемого явления во времени. При этом, статистика может включать и сугубо собственные задачи и методы анализа, вытекающие из сущности исследуемых ею явлений, процессов и объектов. В настоящей книге мы будем рассматривать ту часть аппарата математической статистики, которая наиболее часто используется в статистиках социально-общественных и экономических наук, составляя базу их статистического анализа (*т. н.* ***"Теорию общей статистики"***).

Предложенный нами подход к понятию ***"статистика"*** во многом позволяет объяснить серьезные принципиальные различия между т.н. *"советской"* и *"буржуазной"* статистиками, ибо ничто так не подвластно влиянию идеологии, как *методика* и *методология* исследования,

особенно в общественных явлениях. Именно поэтому официальная советская доктрина полагала, что *"базой для теоретического объяснения статистических категорий и понятий является марксистско-ленинская философия и политэкономия"* [96]. Такой партийный волюнтаризм позволял в нужном свете толковать любые результаты статистического анализа либо применять подходящие методики отбора данных, их классификации, группировки и анализа. В таких строго формализованных науках как теория вероятностей и математическая статистика сделать это значительно сложнее как по причине недостаточной математической грамотности *"ученых-общественников"*, так и высокого уровня абстракции предметов. Следуя вышесказанному, под ***"статистикой"*** в общем случае мы будем понимать методику и методологию применения методов математической статистики (*как прямых, так и модифицированных для конкретных приложений*), а также собственных методов и приемов для анализа массовых социально-общественных и экономических явлений с учетом их внутренней специфики.

## 1.2. Краткий экскурс по истории статистики

Статистика и учет возникли в глубокой древности, за тысячелетия до н.э. Зарождение практической статистики отмечается уже в Китае в 22 в. до н.э. Кроме Китая, практическая статистика находит применение в Египте, Персии и Римской империи. Особенно развитой была своеобразная статистика империи инков. Наиболее характерным является развитие учета и статистики в древнем Риме. Налоговая политика Рима требовала четкого учета дохода его граждан. Имущественные цензы (*учеты*), охватывая всю территорию Римского государства, давали возможность вести подробный хозяйственный учет. Первые учетные операции эпохи феодализма относятся к концу 9 в. Из истории известно об инвентаризации Карла Великого всех королевских имений и учета населения, способного носить оружие. Церковные и феодальные хозяйства вели учет земельных угодий, имущества, скота, доходов и др., закладывая основы (*хотя и примитивные*) статистического анализа.

Серьезным стимулом для развития учета и статистики послужило развитие торговых и международных отношений, а также товарно-денежных отношений в период раннего капитализма. Торговые книги становятся вещественным доказательством в суде; в Италии с 14 в. наряду с простой вводится двойная бухгалтерия, в которой операции фиксируются дважды – в дебете и кредите. Это стало возможным, ибо в качестве дебиторов и кредиторов наряду с людьми стали фигурировать и предметы (*товары*, *оборудование*, *недвижимость и др*.).

В последней трети 17 в. идея использования статистики в социально-экономических исследованиях возникла у английского экономиста У. Петти, пытавшегося на основе числового расчета дать ответы на экономические и социальные вопросы. Петти высказал ряд интересных мыслей по организации статистического наблюдения, широко используя категорию средних. Современник У. Петти Дж. Граунт, анализируя данные о естественном движении населения, обнаружил в них интересные *массовые количественные* закономерности. Последователи Петти и Граунта создали направление статистического анализа, названное *политической арифметикой*. В дальнейшем под термином ***"статанализ"*** будем понимать любое статистическое исследование в целом.

Одновременно с политической арифметикой в Германии во второй половине 17 в. возникло *описательное* направление, основанное Г. Конрингом. Представители этого направления изобретали различные системы описания государства. Впоследствии Г. Ахенваль назвал это описательное направление ***статистикой***, впервые употребив этот термин почти в его современном значении. Со временем описательное направление в статистике лишалось своих сторонников и к середине 19 в. потеряло свое значение. Заслугой этого направления

является создание и введение в научный оборот таблиц и графиков, а также достаточно четкое определение статистики. В этот *период* развитие статистики было связано с прогрессом математики и, особенно, теории вероятностей в трудах П. Ферма, Б. Паскаля, Х. Гюйгенса, Я. Бернулли, П. Лапласа, К. Гаусса, А. Лежандра, Г. Лейбница и др.

В результате массовых статистических наблюдений были обнаружены простейшие типы устойчивости статистических показателей. Главными идеологами и вдохновителями *теории устойчивости* были В. Лексис и А. Кетле. К заслугам Кетле следует отнести открытие им ряда закономерностей массовых явлений. Однако уже во второй половине 19 в. было обнаружено, что статистические показатели отражают изменения экономического цикла. Устойчивость статистических показателей имеет место только в течение непродолжительного периода жизни общества при относительном постоянстве *социально-экономических* условий. Например, постоянство во времени числа рождений, смертей, самоубийств, преступлений и др., приходящихся на тысячу человек населения в конкретной стране, резко нарушается во времена серьезных кризисов. Примером этого может служить сегодняшняя Эстония, в которой первый показатель уменьшился, а остальные резко возросли. Подобная картина имеет место и для других республик бывшего СССР.

В. Лексис предложил математический аппарат для измерения устойчивости, его идеи позднее были положены в основу *дисперсионного анализа*. А. Кетле много занимался вопросами статистической практики не только на родине в Бельгии, возглавляя национальную статистику, но и в масштабе всей Европы. Он был организатором первых международных статистических конгрессов, сыгравших большую роль в развитии практической и теоретической статистик. С полным основанием Кетле можно считать родоначальником современной статистики. Им были выделены средние величины в качестве основного приема статистического анализа. Подробнее о роли В. Лексиса и А. Кетле в деле развития статистики можно найти в [84].

*Математическое* направление в статистике, развитое англичанами К. Пирсоном, В. Госсетом, Р. Фишером и др., расширило идеи А. Кетле. Так, К. Пирсон существенно усовершенствовал теорию корреляции и предложил $\chi$-квадрат критерий для проверки гипотез. В. Госсет разработал теорию малой выборки, имеющую большое научное и практическое значение. Его статьи под псевдонимом *"Стьюдент"* публиковались в журнале ***"Биометрика"***. Значительный вклад в развитие математической статистики и теории дисперсионного анализа внес Р. Фишер.

В 19 в. происходит стремительное развитие статистической практики и хотя этот процесс протекал неравномерно во времени и в отдельных странах, он обладал рядом общих черт. Так, начали появляться государственные статистические службы (*Швеция*, *Франция*, *Бельгия*). К концу столетия государственные статистические службы были сформированы почти во всех развитых странах. Наибольший прогресс ощущался в области статистики населения. К концу 19 в. статистический учет всех важнейших областей общественной жизни проводился уже систематически. В организации прикладной статистики большую роль сыграли А. Кетле и проводимые по его инициативе *международные* статистические конгрессы. В тот же период статистика пополнилась новыми *методологическими* идеями, зародились *теории устойчивости* В. Лексиса (*впоследствии названная* ***теорией дисперсии***), ***теории корреляции*** и ***регрессии*** Гальтона. Математические обобщения теории устойчивости были сделаны А.А. Чупровым и А.А. Марковым. Работы по корреляционному и регрессионному анализам заложили основы знаменитой английской биометрической школы.

Таким образом, в 19 веке окончательно определились основные черты статистического метода (*статанализа*): массовое наблюдение, обобщение данных, их анализ и интерпретация

полученных результатов. Мощное развитие теории статистики 19 в. берет начало в трудах В. Лексиса об устойчивости статистических рядов. От изучения общественных явлений статистика благодаря трудам Гальтона перешла к изучению наследственности и изменчивости, дав импульс развитию ***биометрии***. Приступил к созданию математической школы статистики К. Пирсон. В дальнейшем идеи Гальтона-Пирсона-Лексиса получили теоретическое обобщение в концепции ***стохастической теории статистики***, подготовили быстрое развитие математико-статистических методов, их теоретическое обоснование и применение. Таким образом, работы по *практической* статистике ведутся на протяжении уже 4000 лет, в то время как *теоретическая* статистика насчитывает лишь около 300 лет.

В России учетно-статистические работы также восходят к глубокой истории. Однако по целому ряду причин вплоть до 2-й половины 19 в. они не оказывали сколько-нибудь существенного влияния на мировую статистику в целом. Заинтересованный читатель с этим периодом отечественной статистики может ознакомиться в книге [84]. Однако в период {2-*я половина* 19 *в*. – *начало* 20 *в*.} наметился резкий скачок качественного уровня русской статистики благодаря влиянию математической школы Санкт-Петербургского университета, основанной П. Чебышевым. Им и его учениками А.А. Марковым и А.М. Ляпуновым была создана русская школа теории вероятностей, труды которой получили широкое международное признание. Так, П.Л. Чебышев доказал предельную теорему теории вероятностей, сформулировал в общей форме *закон больщих чисел* (**ЗБЧ**). А.А. Марков определил общие условия, при которых имеет место **ЗБЧ,** и центральную теорему теории вероятностей. Он разработал аппарат описания сложных вероятностных процессов – *цепи Маркова*. В частности, цепи Маркова служат для построения статистических моделей, позволяющих изучать закономерности изменения случайных величин в зависимости от неслучайных величин (*например*, *времени*) при решении конкретных экономических, научных и технических задач. Аппарат цепей Маркова нашел широкое применение в физике и других областях науки и техники. Дальнейшие работы отечественных математиков С.Н. Бернштейна, А.Я. Хинчина, А.Н. Колмогорова, В.И. Романовского и др. по цепям Маркова привели к созданию ***теории случайных процессов,*** имеющей обширные приложения. А.М. Ляпунов предложил новый метод характеристических функций. Обобщая идеи Чебышева и Маркова, он доказал центральную предельную теорему теории вероятностей при значительно более общих условиях, чем его предшественники. Однако математические результаты с трудом проникали в практику статистического анализа, что определялось различными взглядами на теоретическую базу статистики. Пионером применения математических методов в России следует считать В.Я. Буняковского и, в первую очередь, в *демографии*. На дальнейшее развитие статистики в России большое влияние оказали работы Ю. Янсона и А.И. Чупрова. Последующие работы его сына А.А. Чупрова предвосхитили и стимулировали поворот в сторону вероятностного обоснования статистического анализа, способствовали развитию математической статистики и ее логико-философскому обоснованию. Им была разработана теория ***стохастической статистики***, оказавшей большое влияние на развитие статистики как в России, так и за ее пределами.

Традиционное для России политэкономическое направление в статистике в начале 19 века благодаря, в первую очередь, авторитету А.А. Чупрова, стало второстепенным. И только деятельность А.А. Кауфмана в конце 19 в. вернуло ему важное значение в статистике. А. Кауфман определял статистику исключительно как *метод* анализа общественных и иных массовых явлений со своими задачами и приемами исследования. Наши взгляды на предмет статистики во многом отвечают его позиции в данном вопросе.

На рубеже 19-20 веков начали зарождаться *отраслевые* статистики (*сельского хозяйства*, *труда*, *торговли*, *населения и др*.), что говорит о достаточно высоком уровне развития статистики в ее

прикладной части. Здесь же следует отметить работы по прикладной статистике К. Маркса, Ф. Энгельса и В.И. Ленина, позволивших продемонстрировать ее значение для выявления экономических закономерностей. Более того, ленинские принципы организации статистики были положены в основу формирования советской системы государственной статистической службы.

Успехи математики, физики, экономики и других наук, рост практических потребностей в статистическом анализе, стремительный прогресс **ВТ** и средств связи способствовали в 20 веке дальнейшему развитию статистики и ее методов. Повысился теоретический уровень прежних методов и появился ряд новых. Четко обозначились главные направления статистического анализа: статические и динамические исследования. Дальнейшее развитие получила важная для практики *теория выборочного метода* (А. Боули, Е. Нейман, А. Вальд, О. Андерсон и др.). Развитие теорий оценивания и испытания статистических гипотез обязано, в первую очередь, трудам Ф. Эджворта, В. Госсета, Р. Фиёера, Э. Пирсона и др. Новый этап в развитии методов *оценивания* и *испытания* гипотез возник в связи с разработкой *многомерного статистического анализа* (**МСА**), имеющего дело с более, чем одной *характеристикой выборки*. **МСА** имеет дело с оценками, доверительными интервалами, испытаниями гипотез для средней, дисперсии, ковариации, корреляционных характеристик и др. Успехи и прогресс **МСА** связаны с работами Р. Фишера, Г. Хотелинга, С. Уилкса, С. Рао, П. Махаланобиса и др. **МСА** объединяет ряд методов, предназначенных для установления характера и структуры взаимосвязей между компонентами исследуемых многомерных данных. При этом, методы **МСА,** как правило, не опираются на предпосылку о вероятностном характере исследуемых явлений. В социально-экономических исследованиях широкое использование **МСА** началось только в 50-60-х годах прошлого века.

Среди методов **МСА** можно отметить такие ныне широко используемые как: факторный, кластерный и дискриминантный анализы, многомерное шкалирование и др. Углубление представлений о свойствах статистических совокупностей привело к созданию *теории размытых множеств*, основоположником которой является Л. Заде. Дальнейшее развитие методов **МСА** тесно связано с широким использованием класса **ПК,** для которых разработан целый ряд мощных пакетов статистического анализа (***SAS*, *StatGraf*, *Systat*** *и др.*). Развитие метода **МСА** во многом способствовало созданию *новой* научной *дисциплины* **– *эконометрии*,** изучающей количественные стороны экономических явлений средствами математического и статистического анализов [299].

В 1930 г. было создано Международное эконометрическое общество, а с 1933 г. стал выходить журнал "*Эконометрика*". Одним из ведущих направлений эконометрии является построение эконометрических моделей, задача которых состоит в проверке экономических теорий на фактическом материале методами математической статистики. Значительное место в *эконометрии* занимает *теория случайных процессов*, в становление которой большой вклад внес Джон фон Нейман. Данная теория широко использует метод Монте-Карло, основанный на моделировании исследуемого процесса или объекта путем многократных повторений его случайных реализаций. Данный метод весьма эффективен, когда построение аналитической модели сложно либо невозможно, например, при решении многих задач теории массового обслуживания, принятия чисто экономических решений и др.

*Теория статистических игр* (**ТСИ**), базирующаяся на теории решений статистических функций А. Уолде [89], имеет широкие применения, первостепенное значение среди которых имеют две тесно связанные друг с другом области. Это *теория статистического оценивания* и *теория принятия чисто экономических решений*. Одной из первых и до сих пор самых широких областей применения **ТСИ** является статистическая теория оценивания,

составляющая часть математической статистики и применяющаяся на практике при формулировке выводов в статистических исследованиях, т.е. в получении информации о генеральной совокупности как по ней самой, так и по ее выборочным совокупностям. Второй сферой приложений **ТСИ** являются задачи ***микроэкономического*** (*на уровне отдельного предприятия или отрасли*) характера, например, *повышение* качества продукции, определение оптимального уровня товарных запасов, выбор трассы новой линии городского транспорта, выбор участков земли для угодий или застройки и др. Задачи ***макроэкономического*** (*на уровне всего народного хозяйства*) характера являются значительно *более* сложными. Это связано с тем, что принятие *макроэкономических* решений носит часто *разовый* характер, т.е. *байесовская* стратегия, наилучшая при многократном принятии аналогичных решений, теряет здесь свои оптимальные свойства. В качестве примера применения **ТСИ** в макроэкономике можно привести решение задач приобретения лицензий на новую продукцию или технологию, выбора оптимального варианта капиталовложений, разработки месторождений полезных ископаемых, социальной политики и др.

В 20 в. основное внимание на этапе первичной обработки данных стало уделяться не *средним* величинам, а анализу их рядов распределений. При анализе распределения главной задачей является *выравнивание ряда* – испытание гипотезы о законе распределения. В этот же период закончилось формирование ***теории корреляции*** (*уточнены все ее аспекты*, *накоплен богатый опыт применения*, *обеспечена математическая строгость*). В США в 50-70-е годы было создано целое семейство мер связей для неколичественных и неранжируемых переменных: парных, частных и множественных, а также найдены их средние квадратические ошибки. На основе *информационных статистик* К.Э. Шеннона зародился самый общий подход к построению мер связи [33, 104]. Таким образом, зарубежные ученые постепенно развивали методы измерения *связей* в направлении все более точных и реалистичных представлений о явлениях природы и общества.

В 20 веке большое внимание было уделено развитию ***теории индексов*** и анализу динамики ***временных рядов***. При этом, до 20-х годов преобладала *стохастическая* теория индексов, но затем за рубежом возобладала *тестовая* теория индексов, заложенная И. Фишером. Вся зарубежная индексная теория построена в рамках изучения динамики цен и покупательной силы денег. Тест И. Фишера обратимости по факторам составил основу современной аналитической концепции в индексной теории, обеспечил связь методологии индексного анализа с анализируемыми объектами. Он позволил перейти от построения изолированных индексов цен к разработке систем индексов, рассмотрению их *синтетических* и *аналитических* функций. Концентрация производства и усложнение экономических связей в 20 в. поставили задачу выработки методологии анализа динамических рядов и прогнозирования. В процессе исследований в этой сфере были выделены *стационарные* и *нестационарные* временные ряды. Статистический анализ динамики вошел в эконометрию. В изучении динамики зарубежная статистика стремиться к *выявлению* и *измерению* устойчивых компонент: *тренда*, *периодических колебаний*, выделенных *случайных компонент* и др. В этой связи наиболее удачным оказывается метод А. Вальда.

Наконец, дальнейшее развитие производства и усложнение экономических (*как внешних*, *так и внутренних*) связей привело к созданию новых направлений в статистике: статистика на макро- и микроуровнях, сравнительная статистика. Одним из главнейших направлений макростатистики является статистика *национального дохода* (**НД**). Систематические расчеты **НД** правительственными органами впервые начались в СССР (1923), затем в Канаде (1925), Германии (1929), Нидерландах (1931), США (1934) и др. В связи с проблемой **НД** возникла необходимость в научно обоснованных методиках его расчета. С этой целью лауреатом Нобелевской премии В. Леонтьевым (США), активным участником разработки первого

межотраслевого баланса СССР за 1923 - 1924 годы, была предложена матричная модель межотраслевого баланса. Такой подход позволяет предсказывать структурные изменения в народном хозяйстве и особенно успешно применяется там, где государство активно влияет на экономику (Япония, Норвегия и др.). Укрупнение предприятий (*объединения*, *тресты*, *концерны*) и резкий рост объемов хозяйственной деятельности выявили *полную* непригодность традиционных бухгалтерских приемов для решения многих внутрихозяйственных задач. Ключ к их решению могла дать только статистическая методология на уровне отдельного предприятия. *Зарождение* статистики предприятия восходит в Германии к 1911 г. В настоящее время внутрихозяйственная статистика получила большое развитие. Дальнейшее развитие статистики привело к оформлению экономической статистики, ведомственных статистик, статистик населения, труда, медицинской, судебной и др. В свою очередь, ведомственные статистики начали получать более узкую дифференциацию, например, статистика сельского хозяйства включает статистики: растениеводства, животноводства, труда, себестоимости продукции и др. [72, 96]. Сравнительная статистика предлагает методики сравнения статистических данных (*например*, *национального дохода*) в зависимости от времени и места (*сравнения между странами*). Создание международных организаций значительно усилило роль сравнительной статистики и статистики в целом.

Так, *Статистическое бюро Секретариата* ЮНЕСКО занимается подготовкой, анализом, обработкой и распространением статистических данных в области образования, науки и техники, культуры и средств массовой информации. Большую методическую и методологическую роль в развитии международной статистики играют статистические публикации ООН, содержащие систематизированные статистические данные о международных и крупных национальных социально-экономических явлениях. *Международный статистический институт* (**МСИ**) – неправительственная организация, занимающаяся развитием и усовершенствованием статистических методов и их применением в различных областях. Его уставом предусмотрено оказание содействия развитию и совершенствованию статистической методологии в международном масштабе, международной сопоставимости статистических данных, обмену профессиональными статистическими знаниями, издательской деятельности в области статистики. *Статистическая Комиссия* ЭКОСОС ООН охватывает различные вопросы статистики и организации проведения крупнейших статистических наблюдений, разработки методологических основ статистического наблюдения, рекомендации по сбору и анализу данных для ведения международной статистики. Она проводит работу по повышению международной сопоставимости статистической информации и координирует статистическую деятельность специализированных учреждений ООН.

В соответствии с рекомендациями Комиссии работает и *Конференция Европейских Статистиков* (**КЕС**) – постоянный вспомогательный орган ЕЭК ООН, занимающийся изучением тенденций развития статистики в Европе, выработкой рекомендаций по вопросам статистической методологии и организации статистических наблюдений. ***Международная статистика*** исследует уровень, структуру и тенденции социально-экономического развития разных стран. Она основывается на обеспечении международной сравнимости социально-экономических показателей, систематизации статистических данных; включает все отрасли социальной и экономической статистик. Некоторые из основных органов, определяющих развитие международной статистики, рассмотрены выше.

С развитием статистики в СССР в советский период читатель может ознакомиться по книге [84]. Но за исключением отдельных моментов ее функции во многом сводились к учету, а не к анализу и прогнозированию. Но даже учетные данные в целом ряде случаев подвергались фальсификированию. Все это объясняется, в основном, сугубо политическими интересами

правящей тогда *партии-монополиста*. Ведь уже в 30-е годы идеологические установки партии стимулировали официальную доктрину ведущих статистиков страны (А.Я. Боярский, В.Н. Старовский и др.) *статистика приложима только к буржуазному обществу, а в социалистическом плановом хозяйстве для нее нет объекта исследования*. Роль такого "*стимула*" для перспектив развития статистики очевидна. Вместе с тем, совершенно игнорировалась важнейшая роль статистического анализа на различных уровнях для перспективного планирования и прогнозирования. Вероятно, тезис В.И. Ленина "*социализм это, прежде всего, учет*" (т. 35, с. 57, 63; т. 36, с. 263) трактовался буквально. А некоторые учебники социалистических стран по статистике определяли статистику как вид народно-хозяйственного учета ([71], с. 19), внося терминологическую путаницу. Ряд советских учебников прямо указывали на наличие особой "*марксистско-ленинской*" статистики, определяя ее как классовую, партийную науку ([92], с. 10). По нашему мнению, наука, если она вообще таковой является, может быть только правильной или ложной. Именно поэтому деятельность советской (*да и, пожалуй, остальных стран Варшавского договора*) статистики не оказала сколько-нибудь существенного влияния на развитие мировой статистики, но в советском обществе породила не один анекдот. Правда, ряд целый интересных работ по прикладной статистике, выполненных в советский период, получили должное признание и внесли определенный вклад в ее развитие (Б.С. Ястремский, В.С. Немчинов, В.И. Романовский, Д.В. Савинский и др.).

Влияние застойного периода существования статистики в СССР сказывается в ряде случаев и до сих пор (*учебные пособия, вузовские курсы, программы*). Например, вузовская программа [103] *совершенно* не соответствует современным мировым стандартам. Вместе с тем, следует отдать должное успехам советской школы теории вероятностей и математической статистики, обогатившим мировую науку в таких важных разделах как: цепи Маркова, теория случайных процессов, статистика случайных и стационарных процессов в широком смысле (А.М. Яглом, А.Н. Колмогоров, А.А. Боровков, И.И. Гихман, А.В. Скороход, Е.Б. Дынкин, И.А. Ибрагимов, Ю.В. Линник и др.).

Так, работы Ю.В. Линника, В.П. Паламодова и др. способствовали *зарождению* и *становлению* так называемой *аналитической статистики* [106], занимающейся применением современного анализа, в первую очередь, теории *аналитических функций* нескольких переменных к задачам проверки гипотез, оценивания параметров и др. Она имеет дело, в частности, с применением теории функций одной и нескольких комплексных переменных к задачам статистической теории оценивания и проверки гипотез.

Период перестройки предоставил советской статистике новые возможности для исправления негативных аспектов ее развития, но последующий распад СССР завершил 70-летний период советской статистики как мирового явления и перевел ее развитие в русло отдельных самостоятельных государств. В первую очередь, это повлекло за собой пересмотр вопросов организации государственной статистической службы, ее целей и методологии.

## 1.3. Принципы организации государственной статистической службы

Организация статистики в развитых странах во многом определяется их учетно-статистическими традициями, структурой социально-экономического устройства. Можно выделить два основных типа организации статистической службы: *децентрализованная* и *централизованная*. В *первом* случае *сбор* и *обработка* статистической информации производятся соответствующими государственными учреждениями в рамках их компетенции и уже от них данные поступают в специальный статистический орган для сводки и анализа в масштабах всего народного хозяйства. Последний осуществляет лишь методическое руководство статистическими исследованиями с целью обеспечения сводимости и сопоставимости

показателей, поступающих из различных источников. Такая система статистики сложилась во многих странах (США, Великобритания, Франция, Нидерланды, Швеция, Финляндия, Япония и др.).

Так, в США *сбором* статистической *информации* занимаются *многие* агентства. Статистические службы *децентрализованы*, что способствует более *полному* и *своевременному информационному* обслуживанию региональных органов. Координацией статистических программ занимается *Управление стандартов Бюро бюджетов*, которое является координационным центром и по вопросам участия США в международном сотрудничестве по статистике. Сбором первичной статистической информации в США занимаются следующие организации: Бюро бюджетов, министерства сельского хозяйства, торговли, обороны, здравоохранения, образования и благосостояния, *юстиции*, *внутренних дел*, *труда*, *финансов*, Госдепартамент (*в отечественной терминологии*, ***ведомственная*** *статистика*). Большая часть статистических обследований проводится на основе почтовых опросов. В 1980 г. была разработана *единая* информационная система, включающая более 3500 различных *баз данных*, где накапливается и актуализируется разнообразная социально-экономическая информация. Интересное обсуждение вопросов организации статистики за рубежом, включая статистики отдельных отраслей хозяйства, можно найти в [113].

*Централизованный* тип предполагает, что вся статистическая работа концентрируется в специализированных статистических органах. Последние образуют единую иерархическую общегосударственную систему, строящуюся, как правило, по административно-территориальному принципу. При *централизованной* системе создаются наилучшие условия для обеспечения как методологического, так и организационного единства статистического анализа, наиболее эффективного использования трудовых, материальных и финансовых ресурсов, эффективного использования современных средств **ВТ** и связи. В чистом виде статистической службы данного типа не существует (*да и в общем случае это вряд ли возможно*), однако имеет место общая тенденция к повышению уровня ее централизации. При этом достигается более тесное взаимодействие статистики с другими видами социально-экономической информации, с задачами учета и контроля, а также чрезвычайно важными задачами планирования и прогнозирования.

Единая государственная информационно-статистическая служба, базирующаяся на современных средствах **ВТ** и связи, позволяет наряду с соблюдением единых методик и методологий сбора, хранения и обработки данных, производить основную обработку в местах концентрации данных, обеспечивая результатами статистического анализа как заинтересованные органы и организации своего региона, так и вышестоящие уровни службы. При этом, системой распределенных баз данных обеспечивается минимум дублирования информации в службе, достоверность и оперативность. Ближе всего к централизованному типу стояли статистические службы социалистических стран и эта тенденция продолжает сохраняться в новых государствах, образовавшихся из СССР.

Располагая централизованной, в принципе, организацией, статслужба СССР, вместе с тем, находилась под мощным прессом партийных органов со всеми вытекающими из этого негативными последствиями. С этими последствиями первый автор знаком не понаслышке, будучи в периоды 1972-1975 и 1979-1983 г.г. главным инженером и зам. директора по науке Эстонского филиала ВГПТИ ЦСУ СССР. В качестве примера рассмотрим современную организацию государственной службы статистики Эстонской Республики, которая, сохраняя централизованную тенденцию, пытается наиболее активно и успешно войти в европейскую и мировую системы статистики, постепенно приспосабливая к ним свои организационную структуру и методологию.

В общих чертах *государственная служба* статистики Эстонии имеет 3-х *уровневую иерархическую* структуру: *Департамент статистики* (**ДС**) Эстонии, региональные бюро и территориальные отделы. **ДС** формально подчиняется Министерству финансов (*тогда как*, *на наш взгляд*, **ДС** *должен быть независимой государственной структурой*) и включает 9 отделов: макроэкономики, социальной статистики, статистик населения, цен, промышленности, окружающей среды, а также отделы маркетинга и рассылки, развития статистики и информатики. Каждый отдел *курируется* соответствующей *группой экспертов*, тогда как деятельность самого **ДС** курируется Советом по статистике, включающим ведущих специалистов Эстонии по статистике и представителей крупнейших потребителей статистической информации. Совет занимается научными и методологическими вопросами статистики. Отделы, в свою очередь, дифференцированы на сектора. Например, отдел развития статистики состоит из двух секторов: систематизации статистики и методологии, и международного сотрудничества в статистике. Отдел информатики включает сектора: баз данных, регистров, систем классификации и компьютерного обеспечения. Так, в качестве основной *системы управления базами данных* (**СУБД**) выбрана **ADABAS,** в качестве базовой **ВТ –** IBM-совместимые **ПК** и система **RISC 6000,** статистических *пакетов прикладных программ* (**ППП**) **– SAS.** Система **RISC** 6000 обеспечивает координацию функционирования локальных сетей **ДС** и общей сети системы статистики Эстонии, а также работу с основными базами данных на основе **СУБД** ADABAS. В настоящее время наряду с развитием компьютерной сети, охватывающей все органы статистики Эстонии, проводятся работы по созданию регистров, классификации, методике и методологии статистики, базируясь на разработках и рекомендациях Евростата и статистической службы ООН [30].

**ДС** занимается стандартизацией и координацией статистических работ в масштабе всей Эстонии, анализом статистических данных, подготовкой кадров, *методологией* и *методикой* статистики, а также международной статистической деятельностью. Наряду с этим, он обеспечивает правительственные органы и других потребителей *подробной* статистической информацией – фактической и прогнозной. С этой целью **ДС** ежегодно издает 50 статистических отчетов по различным разделам народного хозяйства как на эстонском, так и на английском языках [105]. При этом, часть тиражей предоставляется бесплатно в *правительство*, *крупные* ведомства и международные статистические органы, остальные на коммерческой и заказной основах другим потребителям статистической информации. Тогда как органы 2- и 3-го уровней выполняют задания как для **ДС,** так и для руководящих органов своих регионов. Органы государственной статистики Эстонии проводят свою работу по заранее разрабатываемым годовым планам статистических работ (*утверждаемым Советом по статистике*) под общим *методическим* и методологическим руководством **ДС.**

Помимо органов *государственной* статистики, статистическую работу проводят министерства и департаменты: культуры, социального обеспечения, окружающей среды, налоговый и др. В их составе имеются соответствующие структурные подразделения. Они в пределах своей компетенции собирают и обрабатывают данные, необходимые им для оперативного управления и не собираемые органами государственной статистики. Ведомственная статистика во многих случаях служит источником статистической информации для общегосударственной статистики.

Сбор статистической информации осуществляется в трех основных формах: обязательная статистическая отчетность предприятий и организаций, и специальные статистические анкетирования и опросы. Первая форма, являясь основной, представляет собой совокупность отчетов, содержащих *систему показателей*, *характеризующих* производственно-хозяйственную деятельность за отчетный период. *Отчетность*, как форма наблюдения, имеет три основные особенности: обязательность, юридическая сила и документальная

обоснованность. Эти особенности определяют важные преимущества отчетности как источника статистической информации. Отчетность предприятий и организаций подразделяется на *статистическую* и *бухгалтерскую*, для которой характерна большая направленность на задачи индивидуального контроля. Однако, вторая наряду с первой широко используется как источник данных для статистики. *Анкетирование* позволяет получать дополнительную специальную информацию, тогда как *опросы* служат для повышения оперативности анкетирования. Подробнее о видах и методах сбора статистической информации будет говориться несколько ниже.

## 1.4. Задачи статистики и особенности ее методологии

В период становления рыночных отношений значение учета и статистического анализа резко возрастают по сравнению с социалистической формацией, в сильнейшей степени политизированной. Велика роль статистики в управлении обществом, ибо благодаря только статистической информации управляющие органы могут получать всестороннюю характеристику управляемого объекта, будь то народное хозяйство в целом или отдельные его отрасли или даже предприятия, т.е. реализовывать в системе управления обратную связь. Велико значение статистики для планирования и прогнозирования. Так как в настоящее время планирование многовариантно, потребность его в статистической информации все более и более возрастает. С помощью статистики имеется возможность держать руку на пульсе экономики, во время определяя ее диспропорции.

Статистический анализ количественной стороны общественных явлений проходит *несколько* этапов. На *первом* этапе производится сбор статистических данных о явлении или процессе, подлежащем анализу. Планомерная регистрация существенных признаков элементов статистической совокупности называется ***статистическим наблюдением***. Статистическое наблюдение позволяет охарактеризовать все разнообразие условий и способов проявления изучаемых закономерностей.

На *втором* этапе статистического *анализа собранные* статистические *данные* систематизируются и группируются. Этот этап называют ***сводкой данных*** (*или просто* ***сводкой***). Важнейшим применяемым в ней методом является метод *статистических группировок*, имеющий принципиальное значение потому, что позволяет выделять однородные совокупности, разделять их на группы и подгруппы по существенным признакам и тем самым давать обобщающую характеристику всего объекта. На этом этапе мы переходим от описания отдельных единиц совокупности к описанию выделенных их групп и объекта исследования в целом.

*Третий* этап составляет собственно сам статистический анализ, характеризующийся применением разнообразных методов математической статистики для анализа и обобщения статистических фактов, а также обнаружения закономерностей в изучаемых явлениях. Выводы и сам анализ излагаются, как правило, в текстовой форме с графическими и табличными иллюстрациями. Формы и методы анализа изменяются в зависимости от характера изучаемых явлений и процессов. Необходимость изучения всех аспектов какого-либо процесса требует гибкой системы группировок статистических данных.

***Статистическая методология –*** совокупность приемов, правил и методов статистического исследования социально-экономических явлений: сбор данных, их верификация, обработка, вычисления обобщающих статистических показателей и т.д. Основными здесь являются методы массового наблюдения, группировок и обобщающих показателей. Важнейшей особенностью статистической методологии является конкретность статистического анализа –

примат качественного анализа, связанного с выяснением прежде всего сущности явления с учетом его места и времени развития.

Особенности статистической методологии можно выразить тремя положениями: (**1**) точное измерение и описание массовых данных; (**2**) измерение и анализ дифференциации явлений; (**3**) применение сводных показателей для характеристики явлений и закономерностей их развития. *Первая* особенность настолько свойственна статистике, что ее часто определяли как науку о массовых явлениях, хотя это и не совсем так. В этом плане большое значение для статистической методологии имеет *закон больших чисел* (**ЗБЧ**), суть которого сводится к следующему: в массе индивидуальных явлений общая закономерность проявляется тем полнее и точнее, чем больше их охвачено наблюдением. Поэтому в основе статистического исследования всегда лежит массовое наблюдение фактов. Но **ЗБЧ** не является регулятором процессов, изучаемых статистикой, не объясняет внутреннего механизма процессов формирования закономерностей качественных изменений явлений. Он характеризует лишь одну из форм проявления закономерностей в массовых количественных отношениях.

*Второй* особенностью статистического метода является дифференцированный подход к изучаемым объектам. Совокупность единиц или явлений всегда разбивается на более или менее однородные группы как для того, чтобы установить ее структуру, так и для характеристики основных типов явлений. В распоряжении статистики имеется мощный метод, позволяющий характеризовать и общие закономерности, и тенденции развития, несмотря на то, что они осуществляются путем бесчисленных отклонений от основного правила через многочисленные индивидуальные особенности развития. Это – метод сводных показателей, средних и индексов, с помощью которых в одной или нескольких величинах характеризуется все многообразие массовых явлений. Это составляет *третью* особенность статистического метода.

*Статистическая методология* позволяет исследовать всю совокупность фактов, изобразить процесс в целом, учесть все основные тенденции развития и разнообразие форм явлений, помогая открывать и анализировать причинные зависимости и закономерности массовых явлений и процессов. Здесь имеются в виду *статистические закономерности*, проявляющиеся лишь в большой массе явлений и имеющие теоретической основой **ЗБЧ**. Для обеспечения указанных функций статистика располагает такими средствами, как массовое статистическое наблюдение, система показателей, всесторонне характеризующих хозяйство, групповые и комбинационные таблицы, представляющие результаты статистических группировок, обобщающие показатели, балансовый метод и др.

Таким образом, массовое наблюдение, группировка и сводка его результатов, вычисление и анализ сводных показателей – таковы главные черты метода статистики и в то же время – статистики как метода. Использование современных средств **ВТ** и связи позволяет не только повысить оперативность подготовки и представления статистической информации в *государственные органы*, но и существенно увеличить и усилить ее *аналитические* возможности. Массовое использование средств персональной **ВТ** позволяет кардинально изменить подход к решению данной общезначимой задачи.

## 1.5. Основные понятия и категории статистики

При изучении количественной стороны массовых общественных явлений и процессов статистика использует, кроме ранее упомянутых, еще ряд понятий, относящихся к ее основным категориям: признак, вариация, статистическая совокупность, показатель, система показателей. ***Признаком*** будем называть характерное свойство единиц, объектов, явлений, которое может быть наблюдаемо и измерено. Например, признаками предприятия могут

быть вид выпускаемой продукции, размеры производства, численность персонала и т.д. Выделяются ***качественные*** и ***количественные*** признаки. К первым относятся признаки, отдельные значения которых имеют существенные (*качественные*) различия, для вторых отдельные значения различаются по величине. Если *качественные признаки* могут принимать только одно из двух противоположных значений, то они называются *альтернативными*. В качестве примеров качественных, альтернативных и количественных признаков можно соответственно привести профессию, пол и возраст обследуемых.

По их важности признаки подразделяются на ***существенные*** и ***второстепенные***. Если первые определяют главное содержание явления, то вторые отражают его дополнительные характеристики. Так как статистика в процессе анализа массовых явлений имеет дело, как правило, с существенными признаками, то необходимо уметь отделять их от *второстепенных*. Наряду с этим признаки подразделяются на ***варьирующие*** и ***постоянные***. У первых значения варьируются в *допустимых* пределах, у вторых они постоянны. Статистика имеет дело только с *варьирующимися признаками*. *Диапазон вариации* признака определяется его сутью, например, возраст варьируется от 0 до Н лет. При этом, часто границы диапазона вариации строго определить невозможно, например, вес или рост человека. Отдельное значение признака называется его ***вариантом***. Признаки делятся на ***первичные,*** лежащие в основе программы сбора первичной статистической информации, и ***вторичные,*** характеризуемые в процессе обработки и анализа данных. Например, группировка предприятий по эффективности капитальных вложений основана на вторичных признаках, т.к. для расчета эффективности нужно знать первичные признаки: размер капитальных вложений, прирост прибыли и др.

***Статистическая совокупность*** определяется как множество элементов, объектов или явлений, имеющих один или несколько общих признаков и различающихся между собой по другим признакам. Примерами *совокупностей* являются множества *предприятий* текстильной промышленности, ферм, школ и т.д. В совокупностях, единицы которых взаимосвязаны, можно выделять более однородные совокупности. Совокупность называется ***однородной,*** если самые существенные признаки ее единиц, в основном, одинаковы, в противном случае – ***разнородной***. Совокупность может быть однородной по одним признакам и разнородной по другим. Например, однородная по признаку *"профессия"* совокупность научных работников, разнородна по признакам *"пол"*, *"возраст"*, *"количество публикаций"* и др. Если же состав совокупности в течение определенного временного диапазона остается неизменным, то она называется ***стабильной,*** иначе ***динамической***. Однако такая градация достаточно условна. *Совокупность* называется ***нормальной,*** если распределение численности ее вариантов следует нормальному закону распределения, рассматриваемому несколько ниже.

***Показатель*** является одним из основных понятий статистики вообще и представляет собой обобщенную количественную характеристику социально-экономического явления или процесса в его качественной определенности в условиях конкретного места и времени. Типичными примерами статистических показателей являются: производительность труда, среднегодичное количество публикаций и др. Совокупность показателей, всесторонне отображающих развитие явления, образует ***систему показателей***. ***Сводные*** экономические показатели, относящиеся к сложному комплексу экономических явлений или процессов, иногда называют ***синтетическими*** (*национальный доход*, *национальное богатство и др.*). *Величина* показателя вычисляется, исходя из определенных системы единиц измерения и методологии, что требует серьезных проработок по их обоснованию. Для отображения свойств, *структуры* и *динамики* сложных социально-экономических явлений необходима *система взаимоувязанных* статистических показателей.

***Коэффициент*** – это термин, применяемый в статистике для обозначения некоторых важных относительных величин в специальной области, особенно для отношения разноименных

величин. Чаще всего он применяется к *относительным* величинам интенсивности, например, коэффициенты корреляции, ассоциации, вариации, детерминации и др.

Определяя ***учет*** как способ систематического измерения и анализа массовых общественных явлений с помощью математических методов, в нем можно выделить три основных вида учета: статистический, бухгалтерский и оперативно-технический. *Бухгалтерский* учет фиксирует движение финансовых ресурсов хозяйственных единиц, состав этих ресурсов и их источники. *Оперативно-технический* учет фиксирует собственно производственный процесс (*учет ресурсов*, *сырья*, *готовой продукции*, *производительности труда и др*.). Часто оба эти вида учета называют просто ***учетом***. ***Статистический учет*** – это наиболее сложный, многообразный и способный проникнуть в суть *общественных* процессов вид учета. Общими принципами и методами статистического учета (*мы будем называть* ***анализа***, *что адекватнее отвечает сути дел*а) занимается ***общая теория статистики,*** которая и рассматривается в настоящей книге.

***Общая теория статистики*** – это отрасль статистики, рассматривающая общие категории, понятия, принципы и ее методы. Предметом ее изучения являются наиболее общие свойства количественных отношений социально-экономических явлений. Важнейшими ее разделами являются: методика и методология статистических наблюдений, метод статистических группировок и сводки, статистические показатели, абсолютные и относительные величины, метод средних, вариационные и динамические ряды, связный анализ и индексный метод анализа. Принципы, методы и показатели общей теории статистики используются всеми другими отраслями статистики (*как метода в целом*), в которых они находят свое конкретное качественное наполнение. Однако, так как теория вероятностей является фундаментом математической статистики, на которую опирается общая теория статистики, то в *следующей* главе мы дадим краткое изложение ее элементов. Предполагается, что читатель знаком с основами математического анализа, иначе следующая глава может быть пропущена без особых последствий для последующего понимания. Вместе с тем, знакомство в указанных пределах с основами теории вероятностей позволит читателю более глубоко освоить математическую основу общей теории статистики и различных прикладных статистик.

При компоновке содержания следующей главы авторы исходили из той предпосылки, что статистика более значима для будущих специалистов в области социально-общественных наук по сравнению с теорией вероятностей. Вместе с тем, без навыков владения основными вероятностными понятиями и понимания важнейших ее результатов (*таких*, *как законы больших чисел и центральная предельная теорема*) невозможно говорить о каком-либо серьезном понимании статистических методов. Поэтому следующая вероятностная глава содержит, главным образом, тот минимальный базис, который потребуется для усвоения основной статистической части книги. Мы сознательно избегаем тяжелых в техническом отношении доказательств, но оставляем доказательства, имеющие весьма прозрачный вероятностный смысл, способствующие усвоению материала. Подчеркивается особая значимость *предельных* теорем. При отборе статистических методов, которые включены в курс, основное внимание уделяется их универсальности. Подробно излагаются идеи, лежащие в основе тех или иных статистических приемов, в надежде, что это окажется более полезным для дальнейшего образования, чем обилие технических деталей и, порой, сложных математических выкладок.

# Глава 2.
# Элементы теории вероятностей

***Теория вероятностей*** представляет собой раздел математики, изучающий закономерности случайных явлений. Зарождение теории вероятностей как самостоятельной дисциплины восходит к середине 17 в. в работах Б. Паскаля, П. Ферма и Х. Гюйгенса. Доказательство в 1713 г. Я. Бернулли теоремы, ныне носящей его имя, послужило началом возникновения большой группы теорем, именуемых в общем *законом больших чисел* (**ЗБЧ**), и серьезным стимулом дальнейшего развития теории. В 18 в. и начале 19 в. *теория вероятностей* получила *дальнейшее* развитие в трудах А. Муавра, П. Лапласа, К. Гаусса и С. Пуассона. Наиболее плодотворный период в превращении теории вероятностей в стройную математическую науку связан с трудами выдающихся русских математиков П.Л. Чебышева, А.А. Маркова и А.М. Ляпунова, существенно расширивших результаты предыдущих исследователей. В 18-19 веках и *особенно* в 20 в. теория вероятностей находит весьма широкое применение почти во всех областях естествознания и техники, формируя прикладную науку – ***математическую статистику*.**

Современное развитие теории вероятностей и математической статистики формировалось и формируется в результате международного сотрудничества очень большого числа ученых из разных стран. Русская и советская школы теории вероятностей широко известны в мире трудами С.Н. Бернштейна, А. Н. Колмогорова, А.Я. Хинчина, Б.В. Гнеденко, Ю.В. Линника, В.И. Романовского, В.С. Пугачева и др. В частности, А. Н. Колмогоровым дан теоретико-множественный подход к построению теории вероятностей. Сильные результаты получены по цепям Маркова, теории случайных процессов, математической статистике и другим направлениям [147-152]. Некоторые исторические детали в этом направлении можно найти в разделе **1.2** настоящей книги. В 70-80-х г.к. *большое* развитие получили работы по *исследованию стохастических моделей вычислимости* (***машины Тьюринга*** – *модель последовательных вычислений* и ***однородные структуры*** – *модель параллельных вычислений*). Так, например, наши результаты [4, 12, 23, 153] по обобщению однородных структур на стохастический случай открывают не только новые возможности для приложений, но и предлагают для исследования интересные математические объекты стохастической или квазистохастической природы. Таким образом, современные теория вероятностей и математическая статистика – весьма обширные научные разделы математики, поэтому в *настоящей главе* будут рассмотрены только самые *необходимые* сведения по ним, используемые прямо либо косвенно в дальнейшем изложении.

## 2.1. Классическое понятие вероятности и комбинаторика

Одним из основных понятий теории вероятностей является понятие ***случайного события*** (*или просто* ***события***). ***Событием*** называется всякий факт, который в *результате опыта* может иметь либо не иметь места. С *событиями* связываются *некоторые числа*, характеризующие степень их объективной возможности, называемые ***вероятностями событий***. Существует несколько подходов к понятию "***вероятность*":** классический, статистический и теоретико-множественный. *Статистический* подход основан на понятии частоты события в длинной серии экспериментов. *Теоретико-множественный подход* является

*аксиоматическим* и опирается на элементарные понятия теории множеств. В настоящее время аксиоматический подход является основным подходом в теории вероятностей. Однако, мы будем базироваться на эквивалентном ему *классическом* понятии вероятности, состоящем в следующем.

***Вероятность*** **p(b)** события **b** равна отношению числа **n(b)** благоприятных ему случаев к числу **n** всех возможных случаев в ряду случайных экспериментов опыта, т.е. **p(b) = n(b)/n.** В свете этого определения *статистическое* понятие вероятности принимает следующий весьма простой вид:

$$\mathbf{p(b)} = \lim_{\mathbf{n}\to\infty} \frac{\mathbf{n(b)}}{\mathbf{n}} \qquad (1)$$

Тогда как *теоретико-множественное* понятие вероятности по Колмогорову определяется как: каждому случайному событию **b** сопоставляется некоторое действительное число **p(b)** такое, что $\mathbf{0 \le p(b) \le 1}$. При этом, вероятность *достоверного события* (***И*** - *истина*) равна **1,** а *невозможного* события (**F** – *ложь*) **– 0.**

Формула (1) позволяет во многих случаях непосредственно вычислять вероятности событий. Например, в *урне* находится **m** белых и **n** черных шаров ($\mathbf{m \ge 2}$). Из урны вынимают два шара. Какова вероятность события **b =** {*два белых шара*}? Так как событию **b** благоприятствует ровно **x** случаев при общем числе случаев **y,** то легко вычисляем искомую вероятность:

$$\mathbf{x = C_m^2;} \quad \mathbf{y = C_{n+m}^2;} \quad \mathbf{p(b) = \frac{C_m^2}{C_{n+m}^2} = \frac{m(m-1)}{(m+n)(m+n-1)}}$$

Для значений **m**=15 и **n**=8 на основе выведенной формулы получаем значение вероятности **p(b)** наступления события **b: p(b) =** (15*14)/(23*22) = 0.415. Как правило, такому прямому подсчету поддаются вычисления вероятностей для конечных множеств событий, используя комбинаторные методы, и некоторых типов *геометрических* вероятностей, использующих геометрические соображения.

Например, плоскость разбита квадратной сеткой (*со стороной* **4R**) и в ее узлах расставлены стержни. Сверху случайным образом на плоскость перпендикулярно ей бросается кольцо радиусом **R.** Требуется определить вероятность события **A =** {*попадание кольца на стержень*}. В основу решения положим следующие геометрические соображения. Благоприятными для этого (*относительно*, *не нарушая общности*, *произвольного стержня*) события случаями являются те, в которых центр бросаемого кольца попадает в круг радиуса **R** с центром **O** в основании данного стержня. В качестве множества всех *возможных событий* достаточно выбрать квадрат **ABCD** со стороной размера **2R** с тем же **O**-центром (рис. **2**).

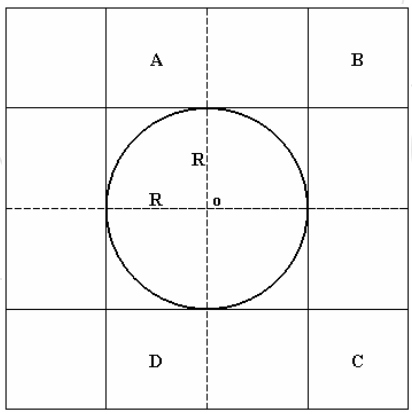



**Рис. 2. Пример вычисления геометрической вероятности**

Таким образом, вероятность **P(A)** наступления события **A** определяется по формуле:

$$P(A) = \pi R^2/4R^2 = \pi/4 \approx 0.785$$

Приведем еще один интересный пример на вычисление геометрической вероятности. На отрезке **[0, m]** наугад ставятся две точки **X** и **Y.** Найти вероятность **P(A)** того, что расстояние **L** между ними будет не больше, чем **d < m.** Геометрически условие **|X-Y| ≤ d** эквивалентно попаданию точки (**X, Y**) в заштрихованную область $S_a$ квадрата размером (**m**x**m**), ограниченную отрезками его сторон и прямыми вида **Y = X + d** и **Y = X - d** (рис. **3, a**).

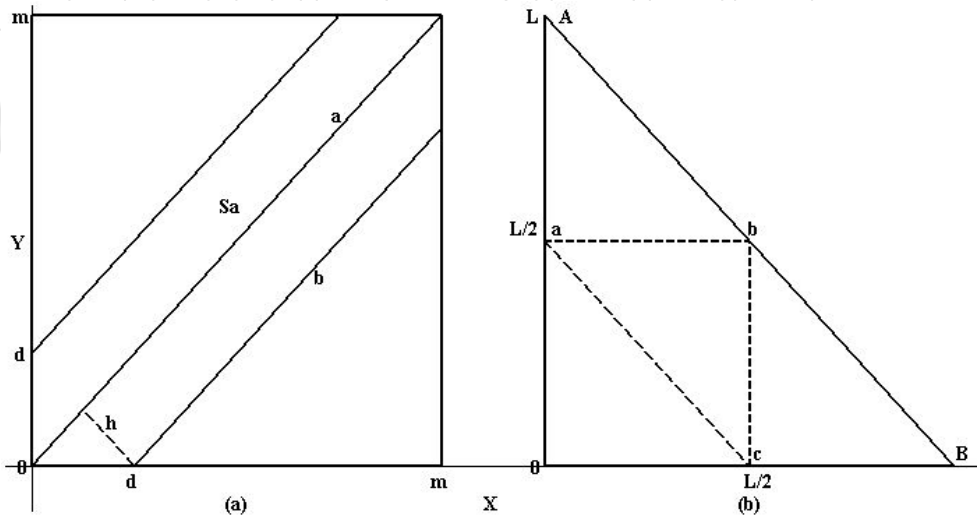



**Рис. 3. Два примера вычисления геометрической вероятности**

Из простых геометрических соображений легко вычисляется площадь $S_a$-области, а именно:

$$h = \sqrt{2}d/2,\ a = \sqrt{2}m,\ b = \sqrt{2}(m-d),\ Sa = (a+b)*h = (2m-d)*d$$

Тогда, вероятность **P(A)** наступления искомого события **A** вычисляется следующим образом:

$$P(A) = Sa/m^2 = (2*m - d)*d/m^2$$

Если в геометрической вероятности пространство элементарных событий не плоское, а трехмерное, то в качестве множеств *случаев*, *благоприятствующих событию* **A,** и всех *возможных* случаев рассматриваются *объемы* соответствующих *пространственных* фигур. Много полезных примеров может быть рассмотрено относительно трехмерной геометрической вероятности, однако это выходит за рамки настоящей книги. *Геометрическая вероятность* была введена, чтобы преодолеть недостаток классического определения вероятности, состоящий в том, что оно неприменимо к испытаниям с бесконечным числом исходов. При таком подходе было предложено приписать событию **A** в качестве вероятности отношение **P(A) = S(A)/S(D)** площади, занятой благоприятными для **A** исходами, к площади *всей* области **D** (*в случае прямой или пространства вместо площади следует рассматривать длину или объем соответственно*). Однако на этом пути появилось немало парадоксов, с которыми читатель может ознакомиться в цитируемой литературе.

Между тем, основными в теории вероятностей являются не прямые, а косвенные методы *вычисления* вероятностей. Для этого, прежде всего, нужно уметь выражать интересующие нас события через другие. Этим целям служит так называемая ***алгебра событий,*** определенная над конкретным пространством элементарных событий. Для дальнейшего изложения определим операции над событиями в графическом виде (рис. **4**), что позволит существенно упростить его понимание.

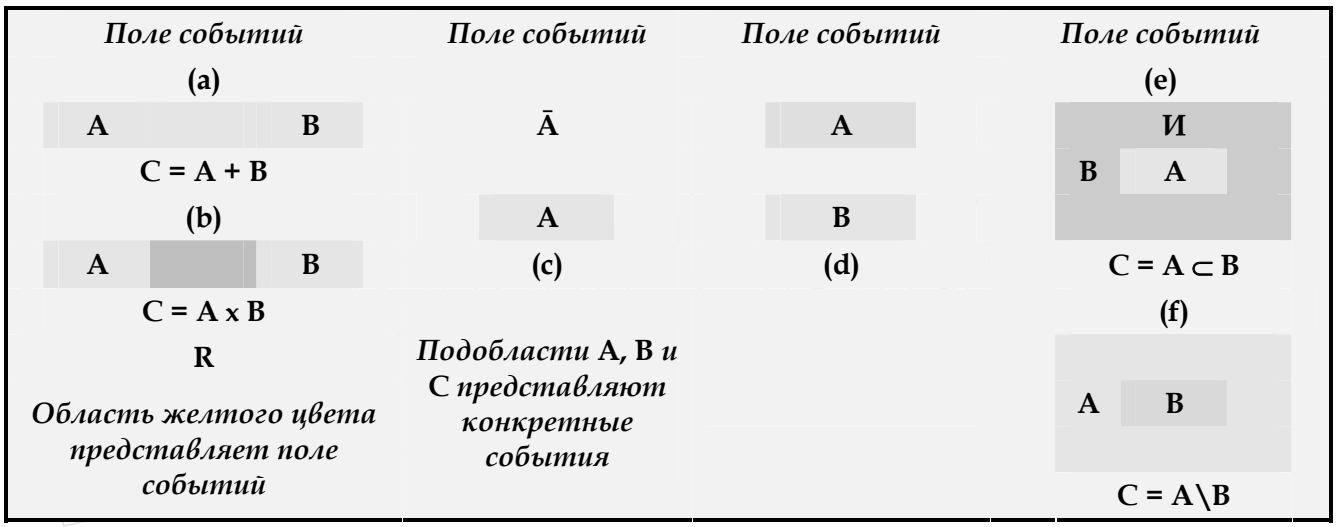



**Рис. 4. Геометрическая иллюстрация алгебры событий**

На рис. **4** поле **R** совокупности всех случайных событий опыта изображено множеством всех точек квадрата (**a**), а отдельные события – в виде точек и областей, лежащих внутри него. Тогда сумма событий **A** и **B** (**C=A+B** или **C=A∪B**), когда произошло по крайней мере одно из этих событий, изображена на рис. **4** (**a**), а их произведение (**C = AxB** или **C = A∩B**), когда оба события произошли одновременно – на рис. **4** (**b**). Операция (**A ⊂ B**) "*событие* **A** *влечет за собой событие* **B**" изображена на рис. **4** (**e**), а ***достоверное*** событие **И** и противоположное к **A** событие **Ā** изображены соответственно на рис. **4** (**c**). ***Невозможное*** событие Л (*ложь*) находится вне поля событий опыта (рис. **4, a**). ***Несовместимые,*** т.е. взаимно исключающие друг друга события **A** и **B** (**AxB = Л** или **A∩B = Ø,** *где* **Ø** – *пустое множество*), изображены на рис. **4** (**d**), а разность событий (**C=A\B**), когда в **C** входят все элементы **A,** не входящие в **B** – на рис. **4** (**f**). Так как события в нашем *теоретико-множественном изложении* представляют собой *множества*, то действия с ними определяются аналогично соответствующим операциям с множествами.

При сделанных предположжениях ***алгебра событий*** принимает следующий простой вид. Вероятности *достоверного* и *невозможного* события соответственно равны **P(И) = 1** и **P(Л) = 0.** Вероятность события **Ā,** *противоположного событию* **A,** равна **P(Ā)=1-P(A).** Это с очевидностью вытекает из их определения.

***Правило сложения вероятностей.*** Пусть теперь $G = \{A_1, A_2, ..., A_n\}$ – конечное множество случайных событий. Тогда вероятность наступления по крайней мере одного из них при условии их попарной несовместимости есть:

$$P(\sum_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j); \quad A_k \cap A_j = empty\ set; \quad k \neq j \qquad (2)$$

Формула (2) обобщается на *бесконечное счетное* множество *попарно несовместимых* случайных событий.

***Правило умножения вероятностей.*** Далее два случайных события **A** и **B** будем называть *независимыми*, если появление одного из них не меняет вероятности появления другого. Вероятность одновременного наступления попарно независимых случайных событий $A_k$ (**k = 1 .. n**) определяется следующим образом:

$$P(\prod_{j=1}^{n} A_j) = \prod_{j=1}^{n} A_j; \quad A_j - mutually\ independet\ events \qquad (3)$$

Формула (3) обобщается на *бесконечное счетное* множество *попарно* независимых случайных (*mutually independent events*) событий.

***Правило условной вероятности. Условной вероятностью*** события **A** при наличии события **B** [*обозначение:* **P(A|B)**] будем называть вероятность события **A,** при условии, что событие **B** произошло. Условие, состоящее в наступлении события **B,** равносильно изменению условий опыта, когда из всех элементарных событий поля **A** (рис. **4**) остаются только те, которые благоприятны событию **B.** Из этого следует, что вместо поля **A** рассматривается новое поле **AB** (рис. **4, b**), соответствующее событию **B.** Тогда область **AB,** соответствующая пересечению **A** и **B,** благоприятна событию **A** при наличии события **B.** Вероятность произведения двух событий **A** и **B** равна вероятности одного из них, умноженной на условную вероятность другого при наличии первого события, а именно:

$$\mathbf{P(A*B) = P(A)*P(B\,|\,A) = P(B)*P(A\,|\,B)} \qquad (4)$$

Таким образом, *условная вероятность* вычисляется по следующим двум формулам:

$$\mathbf{P(B\,|\,A) = P(A*B)/P(A)} \quad \text{или} \quad \mathbf{P(A\,|\,B) = P(A*B)/P(B)} \qquad (5)$$

Правило (4) обобщается на произвольное конечное или счетное бесконечное множество случайных событий, а именно:

$$\mathbf{P\left(\prod_{j}^{n} A_j\right) = P(A_1)\,P(A_2\,|\,A_1)P(A_3\,|\,A_1{*}A_2)\,\ldots\,P(A_n\,|\,A_1{*}\cdots{*}A_{n-1})} \qquad (6)$$

Очевидно, что для независимых событий **A** и **B** будут справедливы следующие соотношения: **P(A|B)=P(A)** и **P(B|A)=P(B),** а формула (6) переходит в формулу (3). Предположим теперь, что каждое событие **B** происходит только одновременно с каким-нибудь из событий $\mathbf{H_j}$ ($\mathbf{j=1\,..\,n}$) (*такие события часто называются* ***гипотезами***), причем события $\mathbf{H_j}$ попарно несовместимы и образуют полную группу событий, т.е. $\mathbf{P(\sum_j H_j]) = 1}$. Для случайного события **B** получаем, используя формулы (2) и (3), следующее важное соотношение:

$$\mathbf{P(B) = P(H_1)*P(B\,|\,H_1) + \ldots + P(H_n)*P(B|H_n) = \sum_j P(H_j)*P(B|H_j)} \qquad (7)$$

Формула (7) называется ***формулой полной вероятности***. Если до опыта вероятности гипотез $\mathbf{H_j}$ были равны $\mathbf{P(H_j)}$, а в результате опыта произошло событие **B,** то уже новые (***условные***) вероятности гипотез вычисляются по следующей важной формуле (*широко известной как формула или теорема Байеса*):

$$\mathbf{P(H_j\,|\,B) = \frac{P(H_j)*P(B|H_j)}{\sum_j P(H_j)*P(B|H_j)} \quad (j = 1, 2, \ldots, n)} \qquad (8)$$

При этом, доопытные (*первоначальные*) вероятности гипотез $\mathbf{P(H_j)}$ называются ***априорными,*** а *послеопытные* $\mathbf{P(H_j\,|\,B)}$ ($\mathbf{j = 1..n}$) *апостериорными*. Формула (8) дает возможность *пересматривать* вероятности гипотез с учетом результатов опыта. Если после опыта, давшего событие **B,** проводится еще один опыт, в результате которого может произойти или не произойти событие **A,** то вероятность (*условная*) этого последнего события вычисляется по формуле (7) полной вероятности, в которую подставлены не прежние вероятности гипотез $\mathbf{P(H_j)}$, а новые апостериорные вероятности $\mathbf{P(H_j\,|\,B)}$ ($\mathbf{j = 1\,..\,n}$), давая следующую формулу:

$$\mathbf{P(A\,|\,B) = \sum_{1}^{n} P(H_j\,|\,B)*P(A\,|\,H_j)} \qquad (9)$$

Данную формулу (9) иногда называют формулой для вероятностей *будущих* событий. Для возможности вычисления вероятностей на основе алгебры событий необходимо знание основ комбинаторики, имеющей дело с подсчетом возможных вариантов. Напомним теперь некоторые из основных комбинаторных формул, полезных при решении многих задач:

| | |
|---|---|
| n! | число *перестановок* из n элементов |
| (n-1)! | число *циклических перестановок* из n элементов |
| $\mathbf{n!/\prod_{j=1}^{k} n!}$ | число *перестановок* из n элементов при наличии k *различных* подгрупп $n_j$ одинаковых элементов |
| $\mathbf{A_n^k = n!/(n-k)!}$ | число *размещений* из n элементов по k |
| $\mathbf{C_n^k = n!/k!(n-k)!}$ | число *сочетаний* из n элементов по k |

С другими полезными комбинаторными формулами вычисления различных вариантов, читатель может знакомиться, например, в книгах [62,94] и, практически, в любом *справочнике* по элементарной математике. Кроме того, современные и математические, и статистические пакеты включают функциональные средства обеспечения вычислений с использованием основных комбинаторных формул [7-11, 15, 18, 22, 34, 35, 38, 39, 42, 44, 45, 50, 97-102, 134-144, 156-161]. Детальнее о *программном обеспечении* для решения *вероятностных* и *статистических* задач речь будет идти в последней главе настоящей книги. Теперь, мы представим ряд примеров на использование алгебры событий для вычисления вероятностей. При этом, в квадратных скобках приводятся краткие пояснения и ответы.

**Пример 1.** *Из ящика с* **n** *перенумерованными изделиями наугад вынимается* **k** (**k < n**) *изделий. Какова вероятность события* **A** *того, что их номера образуют последовательность* **{1, 2, ..., k}?** [*Формула* (**1**): $\mathbf{P(C) = 1/A_n^k = (n-k)!/n!}$].

**Пример 2.** *Игральная кость с 6-ю гранями, пронумерованными от 1 до 6, случайным образом бросается один раз. Найти вероятность наступления событий:* **A=**{*четное число очков*}, **B=**{*не менее* 5 *очков*}, **C=**{*не более* 5 *очков*}. [*Формула* (**2**): **P(A) = 1/2, P(B) = 1/3, P(C) = 5/6**].

**Пример 3.** Какова вероятность события **A** того, что при *трехкратном бросании* кости выпадут цифры **3, 4, 5?** [*Формула* (**3**): **P(A) = (1/6)*(1/6)*(1/6) = 1/216 ≈ 0.005**].

**Пример 4.** Какова вероятность события **AV** того, что при *двухкратном бросании* кости выпадет общая сумма **10?** [*Формулы* (**2**) *и* (**3**): **P(AV) = 1/36+1/36+1/36 = 1/12 ≈ 0.083**].

**Пример 5.** Найти вероятность наступления по крайней мере одного из двух совместимых событий **A** и **N?** [*Используя свойство* **P(A) + P(Ã) = 1** *и рис.* **3** (**c**), *представим событие* **A + N** *как сумму трех несовместимых случайных событий вида* **A*Ñ + Ã*N + A*N: P(A + N) = P(A*Ñ + Ã*N + A*N) = P(A*Ñ) + P(Ã*B) + P(A*N) = P(A)+P(N) - P(A*N)**].

**Пример 6.** Система *космического* базирования за один цикл обзора обнаруживает *космический* объект с вероятностью **t**. *Сколько* потребуется *циклов обзора* для его *обнаружения* с вероятностью **t1** не меньшей, чем **R?** [$\mathbf{t1 = R \le 1 - (1 - t)^n}$ *при условии* **ln(1 - R)/ln(1 - t) ≤ n,** *где* **n** – *заданное число циклов обзора*].

**Пример 7.** *Имеются три одинаковые урны, содержащие белые/черные шары соответственно в количествах* **n/3n, 2n/3n** *и* **3n/0.** *Наугад из одной из урн вынимается шар. Найти вероятность* **P** *изъятия белого шара* [*Используем формулу* (**6**): **P = (1/3)*(1/4 + 2/5 + 1) = 11/20 ≈ 0.550**].

**Пример 8.** Имеется **N** *урн*, каждая из которых содержит **n** белых и **m** черных шаров. Из *первой* урны во вторую перекладывается один шар, затем из второй в третью один шар и т.д. После этого из последней урны вынимают один шар. Найти вероятность **p** того, что вынутый шар окажется белым [*Используем формулу* (**6**): **p = n*(n + 1)/(n + m)*(n + m + 1) + n*m/(n + m)*(n + m + 1) = n/(n + m)**].

**Пример 9.** Из чисел **{1, 2, 3, ..., n}** одно за другим выбирают наугад два числа **a** и **b.** Найти вероятность события **A** того, что **(a - b) ≥ m > 0** [*Используется гипотеза* $\mathbf{H_k}$**={a=k=m+1, ..., n};**

$$\mathbf{P(H_k)=\frac{1}{n};\quad P(A|H_k)=\frac{k-m}{n-1};\quad P(A)=\sum_{m+1}^{n}\frac{k-m}{n(n-1)}=\frac{(n-m)(n-m+1)}{2n(n-1)}}\ ]$$

**Пример 10.** Группа состоит из *отличников*, *хороших*, *средних* и *слабых* студентов в количестве соответственно **a, b, c и d** (**a+b+c+d=n**), которые на предстоящем экзамене могут с равной вероятностью получить соответственно следующие оценки: (**5**), (**4, 5**), (**3, 4**) и (**2,3**). Для сдачи экзамена наугад вызывается один студент. Какова вероятность события **A=**{*получение оценки* **4** *или* **5**} [*Используем гипотезы:* **H1=**{*отличник*}, **H2=**{*хороший*}, **H3=**{*средний*}, **H4=**{*слабый*}**; P(H1) = a/n, P(H2) = b/n, P(H3) = c/n, P(H4) = d/n, P(A) = a/n + b/n + a/2n + 0*P(H4) = (2a + 2b + c)/2n**].

**Пример 11.** Имеется **n** урн, содержащих белые и черные шары. Вероятность наступления события {*извлечен белый шар из* **k**-*й урны*} равна $\mathbf{P_k}$**.** Наугад (*с вероятностью* **P=1/n**) выбирается одна урна, из которой извлекается шар. Какова вероятность того, что мы выбрали **k**-ю урну, если шар оказался белым? [*Используем формулу* (**8**): $\mathbf{G_k}$**=**{*выбрана* **k**-*я урна*}, **A=**{*выбран белай шар*}, $\mathbf{P(A\,|\,G_k) = P_k}$**, P(**$\mathbf{G_k}$**) = 1/n;**

$$\mathbf{P(G_k\,|\,A)=\frac{\frac{P_k}{n}}{\sum_j \frac{1}{n}*P_j}=\frac{P_k}{\sum_{j=1}^{n} P_j}}\ ]$$

**Пример 12.** Отрезок длиной **L** произвольно разбивается точками **X** и **Y** на *три* меньших отрезка. Определить вероятность того, что из этих отрезков можно построить треугольник [*Предположим, что отрезки имеют длины соответственно* **X, Y** и **L - (X+Y).** *Очевидно,* **(X+Y) < L;** *известно также, что сумма двух любых сторон треугольника больше третьей, а разность – меньше третьей. Выбираем* **1**-*й квадрант декартовой системы координат* **XOY** *и выделяем треугольную область* **OAB** (рис. **3, b**), *ограниченную осями координат и прямой* **X+Y = L** *– область всевозможных значений пары* (**X, Y**). *На основе следующих условий:* **X+Y > L - (X+Y), X-Y < L - (X+Y)** и **L - (X+Y) + X > Y** *в ней выделяется треугольная подобласть* **abc,** *точки* (**X, Y**) *которой благоприятны построению некоторого треугольника*, т.е.:

$$\mathbf{P=\frac{S_{abc}}{S_{OAB}}=\frac{L^2/8}{L^2/8}=\frac{1}{4}=0.25}\ ]$$

Целый ряд интересных прикладных задач теории вероятностей с примерами их решения может быть найден в книгах [62, 279, 280, 281, 283]. В то время как обсуждение задач теории вероятностей в контексте продвинутой теории статистики может быть найдено в трехтомном издании [285].

## 2.2. Случайные величины и законы их распределения

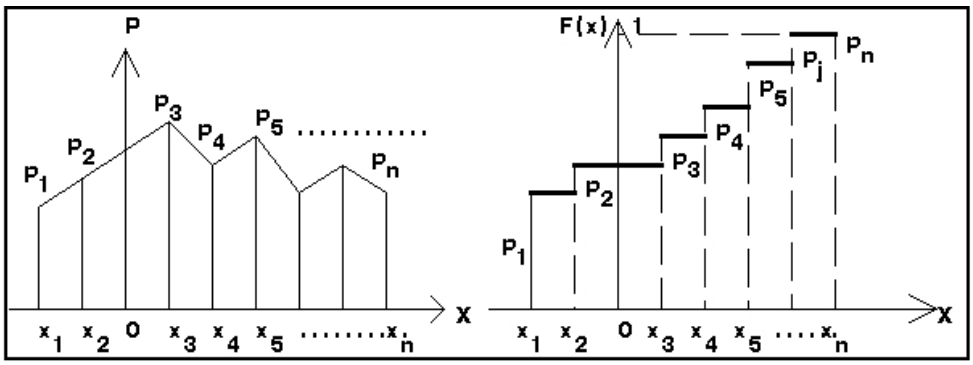
Понятие *случайной величины* (*переменной*) является *одним* из важнейших в *теории вероятностей*. Под ***случайной величиной*** понимается величина, принимающая в результате опыта со

случайным исходом то или иное значение. ***Законом распределения*** случайной величины называется любое правило, позволяющее находить вероятности всевозможных событий, связанных с этой величиной. Например, вероятность попадания ее значения в заданный интервал. Наиболее общей формой закона распределения является ***функция распределения*** **F**, определяющая вероятность того, что случайная величина **X** примет значение меньшее, чем заданное **x: F(x) = P{X < x},** где заглавные буквы будут определять *случайные* величины, а строчные – *неслучайные*. *Функция распределения* **F(x)** для *любой* случайной величины обладает следующими основными свойствами: **F(-∞) = 0, F(+∞) = 1,** а при возрастании **x**-величины функция **F(x)** не убывает. Наиболее простой вид законы распределения имеют у дискретных случайных величин, имеющих конечное или счетное множество значений. Именно данный тип случайных величин характерен для статистических задач. Простейшей формой *закона распределения* дискретной случайной величины **X** является ряд распределения табличного вида, а именно:

| **X:** | $x_1$ | $x_2$ | $x_3$ | ….. | $x_n$ |
|---|---|---|---|---|---|
| | | | | | |
| | $p_1$ | $p_2$ | $p_3$ | ….. | $p_n$ |

$$P_j = P\{X = x_j\}, \quad \sum_j p_j = 1$$

Графическое изображение ряда распределения называется *многоугольником распределения* или *полигоном* (рис. **5, a**). В статистике или в случае использования статистического подхода к понятию вероятности наряду с указанным способом графического представления рядов распределения широко применяются и *гистограммы* (*столбиковые диаграммы*, *секторные и др.*), о которых будет идти речь несколько ниже.




(**a**) (**b**)

**Рис. 5. Графическое представление ряда распределения и функции распределения дискретной случайной переменной**

*Функция вероятностей* **f(x)** является вполне естественным обобщением ряда распределения и определяется как дискретная функция, принимающая значения в точках ее определения, а именно принимает следующий весьма простой вид:

$$\mathbf{f(x)} = \begin{cases} \mathbf{p_j}, & \textit{for } \mathbf{x = x_j} \quad \mathbf{(j = 1,2,3,...)} \\ \mathbf{0}, & \textit{otherwise} \end{cases}$$

С другой стороны, *функция распределения* **F(x)** определяется следующей формулой:

$$\mathbf{F(x)} = \sum_{\mathbf{x_j < X}} \mathbf{f(x_j)} = \mathbf{P(X \le x)} \tag{9}$$

Функция **F(x)** является разрывной, ступенчатой функцией, скачки которой соответствуют возможным значениям случайной величины **X={x1, x2, ..., xj, ...};** между скачками функция **F(x)** сохраняет постоянное значение (рис. **5, b**). В точке разрыва функция **F(x)** равна значению слева (*помечены на рис*. **5, b** *точками*), т.е. она непрерывна слева, тогда как справа может иметь разрыв типа скачка. Тогда вероятность попадания случайной величины **X** на интервал **[a, b)** {*интервал открыт справа, т.е* **a ≤ X < b**} выражается через ее функцию **F(x)** следующей простой формулой:

$$P(a \le X < b) = F(b) - F(a) \qquad (10)$$

В качестве примера случайной величины **X** рассмотрим значение суммы чисел, выпадающих при одновременном бросании двух игральных костей. Используя классическую формулу (**1**), легко составить ее **f(x)-**функцию вероятностей в табличном виде (**a = 35**):

| $x_j$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|
| $f(x_j)$ | 1/a | 2/a | 3/a | 4/a | 5/a | 6/a | 5/a | 4/a | 3/a | 2/a |

Тогда соответствующая ей функция **F(x)** распределения также имеет табличный вид:

| x | <1 | ≤2 | ≤3 | ≤4 | ≤5 | ≤6 | ≤7 | ≤8 | ≤9 | ≤10 | ≤11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| F(x) | 0 | $\frac{1}{a}$ | $\frac{3}{a}$ | $\frac{6}{a}$ | $\frac{10}{a}$ | $\frac{15}{a}$ | $\frac{21}{a}$ | $\frac{26}{a}$ | $\frac{30}{a}$ | $\frac{33}{a}$ | 1 |

Если случайная величина является *непрерывной*, то соответствующая ей *функция распределения* **F(x)** представляется в следующей интегральной форме:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Продифференцировав функцию **F(x),** получаем *плотность* соответствующего распределения **F'(x) = f(x).** Отсюда вытекает справедливость следующего определяющего соотношения:

$$P(a < X \le b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$$

При этом, конкретный вид *закона распределения* зависит от характера вероятностной модели исследуемого явления. В последующих разделах *данной* главы будут рассмотрены некоторые простейшие вероятностные распределения, имеющие наибольшее значение для общей теории статистики, рассматриваемой в настоящей книге.

## 2.3. Характеристики вероятностного распределения

Для характеристики *вероятностного распределения* существует целый ряд числовых величин, из которых здесь мы рассмотрим только наиболее важные для дальнейшего изложения. *Математическое ожидание* (**MO**) (*обозначение* **M[X]=Mx**) *случайной величины* **X** определяется как:

$$M[X] = \sum_{j} x_j * f(x_j) = Mx \qquad (11)$$

Таким образом, **MO** является линейным оператором, что упрощает многие вычисления характеристик вероятностного распределения. Так, **MO** случайной величины **X** из примера раздела **2.2** есть **M[X]= 6.86.** При этом, если сумма (11) не сходится, то **MO** не существует. **MO**

можно рассматривать как стохастическую *среднюю* или *центр рассеяния* значений случайной величины. Для непрерывного случая **MO** принимает следующий интегральный вид:

$$\mathbf{M[X] = \int_{-\infty}^{+\infty} x * f(x)\,dx = Mx} \tag{12}$$

***Дисперсия*** (*вариация*) случайной величины **X** вычисляется по следующей формуле:

$$\mathbf{D[X] = Dx = M[(X - Mx)^2] = M[X^2] - Mx^2} \tag{13}$$

Таким образом, *дисперсия* равна математическому ожиданию квадрата случайной величины минус квадрат ее **MO** и характеризует разброс значений случайной величины относительно ожидаемой ее средней. В статистике рассматривается несколько видов дисперсии, как одной из основных мер вариации. Для *дискретного* и *непрерывного* случаев формулы для дисперсии принимают соответственно следующий вид:

$$\mathbf{Dx = \sum_j (x_j - Mx)^2 * f(x_j);} \qquad \mathbf{Dx = \int_{-\infty}^{+\infty} (x - Mx)^2 * f(x)\,dx} \tag{14}$$

Так, вычисление дисперсии случайной величины **X** из примера раздела **2.2** дает результат **D[X]≈5.52858.** *Среднее квадратическое отклонение* (*или стандарт*) некоторой *случайной* величины **X** вычисляется по следующей весьма простой формуле:

$$\mathbf{y[X] = y_X = \sqrt{Dx}} \tag{15}$$

При вычислении **MO** и *дисперсии* полезны следующие простые правила:

| | | |
|---|---|---|
| **1** | **M[b] = b** | **D[b] = 0** |
| **2** | **M[b*X] = b*M[X]** | **D[X+b] = D[X]** |
| **3** | **M[X+Y] = M[X] + M[Y]** | **D[b*X] = $b^2$*D[X]** |
| **4** | **M[X*Y] = M[X]*M[Y]** | **D[X+Y] = D[X] + D[Y]** |

(16)

где **b** – константа, а правила (**4**) имеют место только при условии независимости случайных величин **X** и **Y.** Доказательства этих правил достаточно просты (*исходя из определений* **MO** *и дисперсии*, *а также свойств конечных сумм и интегралов*) и предлагаются читателю в качестве достаточно полезного упражнения.

***Статистические моменты*** (*или просто* ***моменты***) являются *обобщающими* характеристиками распределения вероятностей и упорядоченных выборок (*вариационных рядов*). ***Момент*** **k**-го ***порядка*** случайной величины **X** относительно произвольного значения **A** определяется следующей формулой (*для дискретного и непрерывного случая соответственно*):

$$\mathbf{M_k = M[(X - A)^k];} \qquad \mathbf{M_k = \int_{-\infty}^{+\infty} (x - A)^k * f(x)\,dx} \tag{17}$$

Из формулы (17) легко следует, при **k=1** и **A=0** имеет место соотношение **M1=M[X],** а при **k=2** и **A=Mx – M2=D[X].** *Момент* **k**-*го порядка* при **A=0** называется *начальным* (**Mk**) и при **A=Mx** – *центральным* (**Mo,k**). Характер *распределения* достаточно хорошо определяется уже *небольшим* числом моментов. На практике часто используются моменты первых четырех порядков. Для упрощения расчетов приведем простые формулы, выражающие центральные моменты через начальные для первых четырех порядков, а именно:

$$\mathbf{Mo,1 = 0,\quad Mo,2 = M2 - M1^2,\quad Mo,3 = M3 - 3*M2*M1 + 2*M1^3}$$

$$\mathbf{Mo,4 = M4 - 4*M3 + 6*M2*M1^2 - 3*M1^4} \tag{18}$$

Читателю рекомендуется вывести формулы (18) и вычислить по ним центральные моменты для случайной величины **X** из примера раздела **2.2** в качестве весьма полезного упражнения. В исследованиях вероятностных распределений используются также моменты произведения и суммы двух и более случайных величин (*смешанные моменты*), но вычисления их требуют специальных приемов [55,62,123,125,156,161,166 181]. Особую роль здесь играют *смешанные центральные моменты второго порядка*, связанные с *коэффициентом корреляции*, который будет рассмотрен несколько ниже. При вычислении числовых характеристик случайных величин часто бывает удобно пользоваться формулой *полных начальных* или *центральных* моментов:

$$\mathbf{Mk} = \sum_{j} \mathbf{P(H_j) * M[X^k \mid H_j]} \qquad \mathbf{Mo,k} = \sum_{j} \mathbf{P(H_j) * M[(X - M_X)^k \mid H_j]}$$

На основе центральных моментов определяется также ряд других важных характеристик вероятностного распределения. Так, например, ***коэффициент эксцесса*,** вычисляемый по следующей простой формуле:

$$\mathbf{E(X)} = \frac{\mathbf{M4}}{\mathbf{M2}^4} - \mathbf{3} \tag{19}$$

Данный коэффициент характеризует степень выделения вершины кривой распределения над всей кривой (*меру крутости распределения*). При этом, *показатель эксцесса* определяется относительно нормальной кривой распределения, для которой имеют место соотношения $\mathbf{M4/(M2)^4 = 3}$ и **E(X) = 0.** При **E(X) > 0** значения случайной величины **X** густо группируются вокруг средней, образуя островершинность кривой распределения; при **E(X) < 0** кривая будет плосковершинной, тогда как при приближении величины **E(X)** к ее нижнему пределу, равному **-2,** двухвершинная кривая распадается на две самостоятельные кривые, что является признаком неоднородности ряда распределения. В качестве полезного примера читателю рекомендуется вычислить величину **E(X)-**коэффициента эксцесса для случая примера из раздела **2.2** настоящей книги.

Для определения *асимметрии* распределения случайной величины **X** относительно вершины кривой используется ***коэффициент асимметрии* A(X),** вычисляемый по простой формуле:

$$\mathbf{A(X)} = \frac{\mathbf{M3}}{\mathbf{M2} * \sqrt{\mathbf{M2}}} \tag{20}$$

Для *симметричного* распределения имеет место соотношение **A(X) = 0;** при **A(X) > 0** имеет место левосторонняя асимметрия; при **A(X) < 0 –** правосторонняя асимметрия. ***Коэффициент вариации*** случайной величины **X** вычисляется по следующей простой формуле:

$$\mathbf{V(X)} = \frac{\sqrt{\mathbf{Dx}}}{\mathbf{Mx}} \tag{21}$$

*Коэффициент вариации* служит для определения принадлежности *случайных* величин *данному* распределению: в случае справедливости соотношения **V(X) ≤ 0.35** принимается тезис о их принадлежности данному распределению. Так, для случая примера из раздела **2.2** получаем оценку **V(X) ≈ 0.343**, говорящую о возможной *принадлежности значений* данному конкретному распределению.

В качестве характеристик положения центра группирования распределения на практике наряду с **MO** и *средней арифметической* иногда используются еще две характеристики: *мода* и *медиана*. ***Модой*** теоретического распределения называется наиболее вероятное значение

случайной величины, а эмпирического распределения – значение, имеющее наибольшую частоту. Если мода для данного распределения единственна, то распределение называют *унимодальным*, в противном случае – *мультимодальным*. Для случая *дискретных* распределений *модами* называются те их значения **xj** случайной **X**-переменной, для которых выполняется следующее определяющее соотношение:

$$\mathbf{P(X = x_j) = \max_k \{p_k\}}$$

***Медианой*** распределения называют такое значение $\mathbf{x_j}$ случайной **X**-величины, для которого имеют место следующие определяющие соотношения: $\mathbf{P\{X < x_j\} \le 1/2}$ и $\mathbf{P\{X \le x_j\} \ge 1/2}$. *Медиана* иногда используется в качестве характеристики центра теоретического или эмпирического распределения. Для рассматриваемых ниже дискретных вероятностных распределений в качестве полезных упражнений рекомендуется вычислить: *коэффициенты эксцесса*, *вариации*, *асимметрии*, *моду* и *медиану*. Более подробно об этих характеристиках вероятностных *распределений* будет идти речь ниже в статистической части настоящей книги.

## 2.4. Основные законы распределения вероятностей

В настоящем разделе рассматриваются *основные законы распределения дискретных* и *непрерывных* случайных величин, представляющие для статистики наибольший интерес. Различают два типа законов распределения: *дифференциальный* и *интегральный*. ***Дифференциальный*** закон **f(x)** дает вероятность отдельного **x**-значения случайной величины **X** (*для непрерывного случая он дает плотность распределения вероятности в точке* **x**). *Геометрическое представление* функции **f(x)** называется *кривой распределения*. ***Интегральный*** закон – универсальная характеристика дискретных и непрерывных случайных величин, определяемая функцией распределения **F(x),** измеряющей вероятность попадания значения случайной величины **X** на интервал **(-∞, x],** открытый слева. Очевидны следующие свойства функций **f(x)** и **F(x)** вероятностного распределения:

$$\mathbf{f(x) = F'_x(x), \quad F(x) = \int_{-\infty}^{x} f(x)\,dx; \quad for \quad f(x) \ge 0}$$

$$\mathbf{F(-\infty) = 0, \quad F(+\infty) = 1, \quad F(x_2) \ge F(x_1); \quad for \quad x_2 \ge x_1}$$

Рассмотрим теперь ряд наиболее важных в практическом отношении *законов распределения*.

***Распределение Бернулли или биномиальное.*** Допустим, вероятность наступления некоторого события **A** в случайном эксперименте есть **p,** а его ненаступления – **q = 1 - p.** Тогда случайная величина **X,** равная *числу наступлений* события **A** в *одном эксперименте*, описывается *функцией* **f(x)** плотности вероятности следующего весьма простого вида: **f(0) = q** и **f(1) = p.** Это есть так называемое *распределение Бернулли*, играющее фундаментальную роль в теории вероятностей и математической статистике, являясь моделью любого случайного эксперимента, исходы которого принадлежат двум взаимно исключающих друг друга классам событий. Легко вычисляются математическое ожидание и дисперсия этого важного распределения (*с точки зрения одного теста Бернулли*):

$$\mathbf{M[X] = 0*q + 1*p = p}$$

$$\mathbf{D[X] = M[X^2] - p^2 = 0*q + 1*p - p^2 = p*(1 - p) = p*q}$$

*Биномиальное распределение* является *моделью* случайных экспериментов, состоящих ровно из **n** независимых однородных испытаний Бернулли. С учетом сделанных выше предположений определяем число **x** *наступления события* **A** при **n** *независимых* экспериментах. Тогда

вероятность наступления события **A** в **n** экспериментах ровно **x** раз, как нетрудно убедиться, вычисляется по следующей простой формуле вида:

$$f(x) = C_n^x * p^x * q^{(n-x)}; \quad C_n^x - \textit{number of combinations} \tag{22}$$

Тогда как функция вероятностей распределения принимает следующий простой вид:

$$F(x) = \sum_{y \le x} C_n^y * p^y * q^{(n-y)} \tag{23}$$

**Пример.** Урна содержит 100 шаров (20 белых и 80 красных). Из нее с возвратом вынимается 3 шара. Определить вероятность наличия среди вынутых шаров 2 белых и 1 красного. Так как вероятность появления белого шара в эксперименте равна **p=0.2,** то используя формулу (20) и простые комбинаторные формулы, легко получаем решение в следующем виде:

$$f(x) = C_3^2 * (0.2)^2 * (0.8)^1 = 0.0096$$

Для **MO** и *дисперсии* биномиального распределения получаем соответственно значения:

$$M[X] = \sum_x x * C_n^x * p^x * q^{(n-x)} = n * p \qquad D[X] = M[X^2] - (n*p)^2 = n * p * q$$

Теперь мы представим доказательства этих выражений на основе полезного продвинутого метода вычисления сумм:

$$Z(p) = \sum_x x C_n^x p^x q^{(n-x)} = \sum_x (x+1) C_n^x p^x q^{(n-x)} - \sum_x C_n^x p^x q^{(n-x)} = \sum_x (x+1) C_n^x p^x q^{(n-x)} - 1$$

$$p + q = 1; \qquad \int_p Z(p)\,dp = p \sum_x C_n^x p^x q^{(n-x)} - p = p(p+q)^n$$

$$\left(\int_p Z(p)\,dp\right)'_p = Z(p) = M[X] = p + q + np(p+q)^{n-1} - 1 = np$$

$$R(p) = \sum_x x^2 C_n^x p^x q^{(n-x)} = \sum_x (x+2)(x+1) C_n^x p^x q^{(n-x)} - 3\sum_x x C_n^x p^x q^{(n-x)} -$$

$$- 2\sum_x C_n^x p^x q^{(n-x)} = \sum_x (x+2)(x+1) C_n^x p^x q^{(n-x)};$$

$$\iint_p R(p)dpdp = \sum_x C_n^x q^{(n-x)} \iint_p (x+2)(x+1) p^x dpdp - \frac{np^3}{2} - p^2 = p^2 \sum_x C_n^x p^x q^{(n-x)} -$$

$$- \frac{np^3}{2} - p^2 = p^2 (p+q)^n - \frac{np^3}{2} - p^2;$$

$$\left[\iint_p R(p)dpdp\right]''_p = R(p) = [2p(p+q)^n + np^2(p+q)^{n-1} - 3np^2/2 - 2p]'_p = np + n^2p^2 - np^2;$$

$$D[X] = R(p) - n^2p^2 = npq \qquad \text{Q.E.D.}$$

Читателю рекомендуется рассмотреть предлагаемые доказательства, которые могут быть полезны при более глубоком ознакомлении с предметом.

***Распределение Пуассона.*** Если число **n** экспериментов при опытах типа Бернулли довольно велико, а вероятность **p** успеха в *каждом* отдельном эксперименте *весьма мала*, то *биномиальное*

распределение можно *аппроксимировать* посредством *распределения Пуассона*. Это *распределение* является предельным для биномиального при следующих условиях: **p -> 0, n -> ∞** и **n*p = a = *const*.** Вывод функции вероятностей **f(x)** для *данного* распределения с использованием весьма несложных преобразований и *асимптотической формулы* Стирлинга для вычисления *значений* (**n!**) при достаточно больших значений **n,** а также в виду соотношений **p = a/n** и **q = 1 - p,** приводится ниже:

$$f(x)=\lim_{n\to\infty} C_n^x p^x q^{(n-x)} = \lim_{n\to\infty} \frac{n!a^x(1-a/n)^n}{x!(n-x)!n^x(1-a/n)^x} = \frac{a^x}{n}\lim_{n\to\infty}(1-a/n)^n(1-a/n)^{-x}\lim_{n\to\infty}\frac{n!}{x!(n-x)!n^x} =$$

$$= \frac{a^x}{x!}\lim_{n\to\infty}\frac{n!}{(n-x)!n^x}.$$ ***Then, using the Stirling formula*** $n!\approx (n/e)^n\sqrt{2pn}$, ***we obtain:***

$$\lim_{n\to\infty}\frac{n!}{(n-x)!n^x} = \lim_{n\to\infty}\frac{n^n e^{n-x}\sqrt{2pn}}{e^n n^x (n-x)^{n-x}\sqrt{2p(n-x)}} = \sqrt{\frac{1}{1-x/n}}\lim_{n\to\infty}\frac{n^{n-x}e^{-x}}{(n-x)^{n-x}} = e^{-x}\lim_{n\to\infty}(1+x/(n-x))^{n-x} = 1$$

$$\textit{Hence,}\quad f(x)=\frac{a^x e^{-a}}{x!}$$

Из данного вывода получаем следующее представление *функции вероятностей распределения Пуассона* **f(x)**:

$$f(x) = \frac{a^x e^{-a}}{x!} \tag{24}$$

где **a** – *параметр* распределения и **e ≈ 2.718281828** – основание натурального логарифма. Тогда Соответствующая функция **F(x)** распределения Пуассона принимает следующий вид:

$$F(X) = e^{-a}\sum_{n\le x}\frac{a^n}{n!} \tag{25}$$

**Пример.** Некоторое изделие производится в массовых количествах. По оценкам известно, что доля *дефектных изделий* составляет величину **p=0.01.** Какова вероятность того, что в *случайной* выборке объемом в **n=100** изделий найдутся три *дефектных*? Используя формулу (24) при **n = 100, p=0.01** и **a=n*p=1,** легко получаем искомое решение: $f(x = 3, a = 1) = e^{-1}/3! \approx 0.06$.

Для *математического ожидания* и *дисперсии* распределения Пуассона получаем идентичные значения, а именно: **M[X] = D[X] = n*p = a,** доказательства которых с использованием весьма несложных преобразований и известной из математического анализа *формулы* представления экспоненциальной функции в виде ряда Маклорена представлены ниже:

$$e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!};\qquad M[X] = e^{-a}\sum_x\frac{x a^x}{x!} = a\,e^{-a}\sum_x\frac{a^{x-1}}{(x-1)!} = a\,e^{-a}e^{-a} = a$$

$$D[X] = M[X]^2 - a^2 = e^{-a}\sum_x\frac{x^2 a^x}{x!} - a^2 = a\,e^{-a}\sum_x\frac{x a^{x-1}}{(x-1)!} - a^2 = a\,e^{-a}\sum_x\frac{(x-1)\,a^{x-1}}{(x-1)!} +$$

$$+ a\,e^{-a}\sum_x\frac{a^{x-1}}{(x-1)!} - a^2 = e^{-a}\sum_x\frac{a^x}{(x-2)!} + a - a^2 = a^2 - a^2 + a = a \qquad \textbf{Q.E.D.}$$

*Распределение Пуассона* является вполне приемлемой вероятностной моделью для описания случайного числа появления определенных событий в фиксированных промежутке времени или области пространства.

***Гипергеометрическое распределение.*** Данное распределение *применяется* в тех случаях, когда выборка элементов делается *без их возврата* в *генеральную* совокупность. Такое распределение, в частности, имеет большое значение для статистического контроля качества продукции. Функция **f(x)** *вероятностей* данного *распределения* является *решением* следующей классической задачи: из урны, содержащей **M** белых и **(N-M)** черных шаров, вынимается по одному шару без возврата. Случайная величина **X** – число белых шаров в выборке из **n** экспериментов. Очевидно, общее число случаев равно $\mathbf{C_N^n}$, а число случаев, благоприятствующих появлению в выборке ровно **x** *белых* и (**n-x**) *черных* шаров, равно следующей величине $\mathbf{C_M^x C_{N-M}^{n-x}}$. Теперь, Используя классическую формулу (1) вероятностей, получаем функцию **f(x)** вероятностей *гипергеометрического распределения* в виде*:*

$$\mathbf{f(x) = \frac{C_M^x C_{N-M}^{n-x}}{C_N^n} \quad (x = 0, 1, 2, ..., a)}$$

Величины **n, M** и **N** называются параметрами *гипергеометрического* распределения.

**Пример.** *И*з 10 студентов семь сдали сессию на 3 и 4, а три – на 5. Наугад выбирается из них пять человек. Найти вероятность того, что среди отобранных окажется два отличника. Эта задача является типичной для *гипергеометрического* распределения при **n=5, M=3** и (**N-M**)**=7**:

$$\mathbf{P(x = 2) = \frac{C_3^2 C_7^3}{C_{10}^5} = \frac{5}{12} \approx 0.417}$$

*Математическое ожидание* и *дисперсия гипергеометрического* распределения вычисляются по формулам соответственно:

$$\mathbf{M[x] = n*p, \quad D[X] = (N-n)*n*p*q/(N-1); \quad q = 1 - M/N}$$

Вывод этих формул предлагается читателю в качестве весьма полезного упражнения.

***Геометрическое распределение***. Данное распределение описывает следующую модель опытов по схеме Бернулли. В результате проведения ряда независимых экспериментов требуется однократно добиться конкретного результата **A,** имеющего вероятность **p.** Здесь в качестве случайной величины **X** выбирается число попыток, требующихся для реализации **A**-события один раз. Нетрудно получить функцию **f(x)** вероятностей такого распределения, а именно:

$$\mathbf{f(x) = p*(1 - p)^x \qquad (x = 0, 1, 2, \ldots)}$$

**Пример.** Производится ряд попыток завести двигатель автомобиля. Каждая попытка завода заканчивается *успешно* независимо от других с вероятностью **p=0.75** и требует **t=5** с. времени. Найти вероятность **P** *успешного* запуска двигателя в течение 30 с. Задача хорошо вписывается в модель *геометрического* распределения, поэтому получаем: **n** = 30/5 = 6 – максимально допустимое число таких попыток; тогда как **q** = (1 - 0.75) = 0.25 – вероятность неуспешной попытки и $\mathbf{p_n = q^6}$ = $(0.25)^6$ = 0.000244 – вероятность незапуска двигателя в установленное время. Тогда вероятность *противоположного* ему события, очевидно, равна $\mathbf{P=1 - p_n =0.999756}$.

**MO** и *дисперсия геометрического* распределения вычисляются соответственно по формулам:

$$\mathbf{M[X] = q/p \qquad D[X] = M[X^2] - (q/p)^2 = q/p^2}$$

Доказательство формулы для **M[X]** с использованием *несложных* преобразований и известной формулы для суммы бесконечной геометрической прогрессии приведено ниже:

$$M[X] = \sum_x xp(1-p)^x = p(\sum_x (x+1)q^x - \sum_x q^x) \qquad x = 0, 1, 2, \ldots, \infty$$

$$M[X] = p(\sum_x (x+1)q^x - 1/(1-q)) \qquad Z(q) = \sum_x (x+1)q^x \qquad \int_q Z(q)dq = q\sum_x q^x = \frac{q}{1-q}$$

$$Z(q) - \left(\frac{q}{1-q}\right)'_q = 1/(1-q^2) = 1/p^2 \qquad M[X] = p(1/p^2 - 1/p) = (1/p - 1) = q/p \quad \textbf{Q.E.D.}$$

Аналогичным методом вычисляется и дисперсия распределения, что может послужить читателю хорошим упражнением. Важность *геометрического* распределения объясняется так называемым свойством отсутствия последействия, а именно: для любых $m, n \geq 0$ имеет место следующее соотношение: $P\{X \geq m + n \mid X \geq m\} = P\{X \geq n\}$.

Наряду с рассмотренными имеется еще целый ряд интересных и полезных дискретных распределений (*Pascal*, *Pearson*, *Erlang*, *Gompertz*, *Frechet*, *Fisher*, *Wishart*, *Student*, *Rayleigh, Poisson*, *Polya*, *von Mises*, *Pareto*, *Laplace*, *Cauchy*, *Dirichlet*, *Weibull*, *Helmert*, *Borel-Tanner и др.*). Например, распределение Пойа (*Polya*) широко используется при моделированиях распространения эпидемий инфекционных заболеваний, а распределение Бореля-Таннера (*Borel-Tanner*) играет весьма большую роль в исследованиях поведения систем массового обслуживания. С рассмотренными, вышеупомянутыми и другими дискретными распределениями, читатель может ознакомиться достаточно подробно в книгах [43,48,52,54,55,63,132,155,161,182,279-281].

В отличие от *дискретного* в *непрерывном* случае множество значений случайной величины **X** несчетно, ибо *непрерывно* заполняет какой-то отрезок оси **X**-ов *декартовой* системы координат. Аналогично дискретному случаю, функция распределения **F(x)** случайной величины **X** определяется как **F(X) = P{X < x},** а функция **f(x) = F'(x)** называется *плотностью вероятности* (*или плотностью распределения*). Вообще говоря, в качестве плотности распределения может быть выбрана любая интегрируемая функция **f(x),** удовлетворяющая лишь двум основным условиям:

$$f(x) \geq 0 \qquad \int_{-\infty}^{+\infty} f(x)\,dx = 1$$

Данные условия непосредственно вытекают из самого определения функции плотности вероятности. Существует целый ряд часто встречающихся на практике непрерывных распределений, но наиболее важным из них является несомненно *нормальное распределение*. Введенное Гауссом, нормальное распределение является непрерывным приближением биномиального распределения и имеет фундаментальное значение, так как при довольно широких предположениях суммы случайных величин с ростом числа слагаемых ведут себя асимптотически нормально. Именно соответствующие условия этого составляют содержание центральной предельной теоремы теории вероятностей.

Основное содержание теории вероятностей наряду с **ЗБЧ** составляет и *центральная предельная теорема* (**ЦПТ**). Суть **ЦПТ** состоит в том, что распределение большого числа (**n**) однородных случайных величин близко к нормальному и при неограниченном увеличении величины **n** стремится к нему. Наиболее простая формулировка **ЦПТ** состоит в следующем: если **X = {X1, X2, X3, ..., Xj, ...}** – *независимые одинаково распределенные* случайные величины с математическим ожиданием **M[X]** и дисперсией **D[X],** то стандартизованная сумма следующего вида:

$$\frac{\sum_{1}^{n} x_j - n*M[X]}{\sqrt{n*D[X]}}$$

в качестве предельного при **n ->** ∞ имеет *нормальное* распределение с параметрами **M[X] = 0** и **D[X] = 1.** В общем виде **ЦПТ** была впервые доказана А.М. Ляпуновым. Данная теорема в некотором роде устанавливает исключительность нормального распределения. Вместе с тем, следует иметь в виду, что нормальное распределение не может быть универсальным. В частности, в социально-экономических приложениях наличие эксцесса и асимметрии часто не позволяет выдвигать гипотезу о нормальном распределении генеральной совокупности. Пример подобного рода, связанный с анализом распределений публикаций **ТТГ,** будет рассмотрен в статистической части книги. На практике нормальное распределение имеет важное значение по целому ряду весьма существенных причин, прежде всего, так как:

- *при экспериментах и наблюдениях многие случайные величины оказываются нормально распределенными;*
- *целый ряд случайных величин являются приближенно нормально распределенными;*
- *случайную величину, не распределенную нормально, обычно можно так преобразовать, чтобы преобразованная величина имела распределение, близкое к нормальному;*
- *нормальное распределение может служить аппроксимацией для многих других важных вероятностных распределений;*
- *при проверке статистических гипотез часто появляются распределения, являющиеся нормальными или, в той или иной мере, близкими к ним.*

Итак, *нормальное распределение* – это непрерывное распределение с плотностью вероятности **f(x),** определяемой следующей интегральной формулой:

$$f(x) = \frac{1}{\sigma\sqrt{2p}} e^{\frac{-(x-M[X])^2}{2\sigma^2}} \qquad D[X] = \sigma^2 \tag{26}$$

Тогда как соответствующая ему функция **F(x)** распределения представляется интегральной формулой следующего вида:

$$F(x) = \frac{1}{\sigma\sqrt{2p}} \int_{-\infty}^{x} e^{\frac{-(x-M[X])^2}{2\sigma^2}} dx \tag{27}$$

Поэтому, вероятность того, что случайная величина **X** примет значение внутри *произвольного* интервала **(a, b],** замкнутого справа, вычисляется как **P(a < X ≤ b) = F(b) - F(a)**. Параметрами нормального распределения являются *математическое ожидание* **Mx** и *дисперсия* **D[X]** (26). На практике часто используется так называемое *стандартное нормальное распределение* (**СНР**), для которого имеют место следующие простые соотношения: **M[X] = 0** и **D[X]=1**. Стандартизация достигается линейной заменой переменных вида **z = (t - M[X])/σ** в формулах (26, 27). Функция плотности **n(z)** и (*кумулятивная*) функция распределения **N(z)** для **СНР** принимают вид:

$$M[X] = 0; \quad \sigma = 1; \quad D[X] = \sigma^2 = 1; \quad z = (t - M[X])/\sigma$$

$$n(z) = \frac{e^{-z^2/2}}{\sqrt{2p}} \qquad N(z) = \frac{1}{\sqrt{2p}} \int_{-\infty}^{z} e^{-t^2/2} dt \tag{28}$$

Графически функции плотности **n(z)** и распределения **N(z)** для стандартного нормального распределения представлены на рис. **6**.

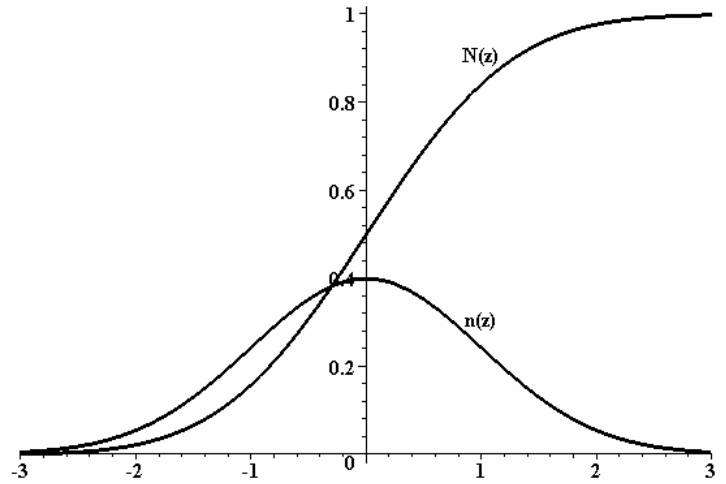



**Рис. 6. Графики функций вероятности n(z) и N(z) для СНР**

Итак, вероятность того, что случайная величина **X** лежит на отрезке **(a, b],** замкнутом справа, вычисляется по следующей формуле:

$$P(a < x \le b) = F(b) - F(a) = N((b - M[X])/\sigma) - N((a - M[X])/\sigma) \quad (29)$$

Вероятности для любых нормально распределенных случайных величин вычисляются на основе формулы (29) и таблицы стандартного нормального распределения, как правило, приводимой в большинстве справочников по математике и книгах по теории вероятностей и математической статистике, а также по общей статистике [48, 49, 51-55, 57, 58, 60-63, 83, 92-94, 116, 132, 155, 156, 166, 182, 186] и др. Кроме того, большинство современных математических и статистических пакетов имеют встроенные средства для представления функций наиболее распространенных распределений (*и дискретныхе*, *и непрерывных*) наряду с вычислением их значений [8, 11, 15, 20, 29, 30, 34, 36, 38, 42, 44, 50, 100-103, 128, 132, 134-144, 165, 169]. Некоторые из этих средств будут рассмотрены в заключительной главе настоящей книги.

Вообще говоря, для значений нормально распределенной случайной величины **X** имеют место следующие приближенные равенства:

$$P(M[X] - \sigma < X \le M[X] + \sigma) \approx 0.68$$

$$P(M[X] - 2\sigma < X \le M[X] + 2\sigma) \approx 0.955$$

т.е. **68%** всех возможных значений любой нормально распределенной случайной величины **X** находятся на интервале **(M[X] - σ, M[X] + σ],** замкнутом *справа*, тогда как **95.5% –** на интервале **(M[X] - 2σ, M[X] + 2σ],** замкнутом *справа*. Нетрудно убедиться, что *мода* и *медиана нормального* распределения совпадают с его *математическим ожиданием* **M[X].**

**Пример.** Случайная величина **X** подчинена *нормальному* закону распределения с параметром **Mx = 0.** Пусть требуется определить, при каком значении дисперсии вероятность попадания случайной величины **X** в *заданный* интервал (**a, b**) при **0 < a < b** (рис. **6, a**) достигает *максимума*? Решение данного примера может быть получено достаточно просто. Основная идея решения сводится к максимизации функции (*вероятность попадания случайной переменной в заданный интервал*) от одной **σ**-переменной (*дисперсия*). Эта проблема легко решается классическими средствами математического анализа, а именно:

$$P(a < X \le b) = F(b/\sigma) - F(a/\sigma) = R(\sigma) = \frac{1}{\sqrt{2p}}\left[\int_0^{b/\sigma} e^{-t^2/2}\,dt - \int_0^{a/\sigma} e^{-t^2/2}\,dt\right]$$

$$R'(\sigma)_\sigma = 0 = \frac{1}{\sqrt{2p}}\left[\frac{-b}{\sigma^2}e^{-b^2/2\sigma^2} - \frac{-a}{\sigma^2}e^{-a^2/2\sigma^2}\right]; \quad \textit{hence}, \quad y = \sqrt{\frac{b^2 - a^2}{2\ln(b/a)}}$$

Мы рекомендуем читателю проверить это решение как достаточно полезное упражнение.

Поскольку в общепринятом обозначении дисперсия обозначается греческой сигмой (**σ**), то в дальнейшем термины *"вариация"*, *"дисперсия"*, и *"сигма* (**σ**)" мы будем часто отождествлять. Нормально распределенная случайная величина **X** с большой вероятностью принимает значения, близкие к своему **MO,** что выражается известным ***правилом сигм*** (*дисперсий*; *одной*, *двух*, *трех и т.д.*), а именно:

$$P\{|X-M[X]| \ge k^*\sigma\}=Pk;\ P1 \approx 0.3173,\ P2 \approx 0.0455,\ P3 \approx 0.0027$$

Чаще всего в практических расчетах используется именно правило ***трех сигм*** (**3σ**).

## 2.5. Основные критериальные распределения

В предыдущем разделе рассматривалась группа распределений, непосредственно связанных с математическими моделями случайных экспериментов. Следующую группу составляют так называемые *критериальные* распределения. Они являются основой статистических критериев, т.е. их функции распределения используются при проверке статистических гипотез. К этой группе относятся и некоторые распределения *первой* группы, например, важное *нормальное* распределение. Здесь мы представим следующие три распределения, играющие важную роль в статистическом анализе: **χ**-*квадрат распределение Хельмерта*, **t**-*распределение Стьюдента* и **F**-*распределение Фишера*.

***Распределение χ-квадрат,*** введенное Ф.Р. Хельмертом, лежит в основе так называемых **χ**-квадрат критериев. Пусть **X={X1, X2, ..., Xn}** – независимые случайные величины, имеющие стандартные нормальные распределения. Сумму квадратов этих величин обозначим через **χ**, а соответствующее ей *распределение* будем называть **χ**-*распределением* (**χ**-*квадрат распределением*) с плотностью распределения следующего вида:

$$\chi = \sum_{j=1}^{n} x_j^2; \qquad f(x) = K_n\, x^{(n-2)/2} e^{-x/2}$$

где **n** – число степеней свободы распределения, а **Kn** – константа, выбираемая так, чтобы для соответствующей функции распределения **F** выполнялось условие **F(∞) = 1**. Именно данное обстоятельство, наряду с рядом других соображений, лежит в основе весьма широкого применения в теории вероятностей и статистике **χ**-квадрат распределения. Вид функции **f(x)** полностью определяется ее **n**-параметром, но уже при **n > 30** **χ**-квадрат распределение переходит в известное нормальное распределение. Например, для четных степеней свободы **n=2*m** получаем **χ**-квадрат распределение с плотностью вероятностей следующего вида:

$$f(x) = K_{2m}\, x^{m-1} e^{-x/2}; \qquad x > 0 \quad (m = 1,2,3,...,30)$$

для которого значения $K_{2m}$, **M[X]** и **D[X]** вычисляются следующим образом. Используя замену **x**-переменной (**x = 2t, dx = 2dt**), мы получаем следующее преобразование:

$$K_{2m}\int_0^{\infty} x^{m-1} e^{-x/2}\,dx = 1 \quad \Rightarrow \quad K_{2m}\frac{1}{2^m}\int_0^{\infty} t^{m-1} e^{-t}\,dt = 1$$

Интегрирование которого дает следующий результат:

$$\int_0^{\infty} t^{m-1} e^{-t} dt = \left[ -e^{-t} \sum_{j=1}^{m-1} t^j \frac{(m-1)!}{j} - (m-1)! e^{-t} \right]_0^{\infty} = (m-1)!$$

Теперь мы можем легко получить следующий результат:

$$K_{2m} = \frac{1}{2^m (m-1)!}$$

В обшем случае, для положительного целого **n** имеет место следующая формула:

$$K_n = \frac{1}{2^{n/2} \Gamma(n/2)}$$

где **Γ** – гамма функция, широко используемая в теории вероятностей и математической статистике [4, 55, 61-63, 83, 93, 96, 125, 166, 186]. Мы можем легко получить выражения для математического ожидания **M[X]** и дисперсии **D[X]** для **χ**-квадрат распределения (*для случая четных степеней свободы* **n = 2*m**), а именно:

$$M[X] = K_{2m} \int_0^{\infty} x^{m+1} e^{-x/2} dx = 2^{m+1} \frac{m! 2^{m+1}}{(m-1)! 2^{m+1}} = 2m = n$$

$$D[X] = K_{2m} \int_0^{\infty} x^{m+1} e^{-x/2} dx - 4m^2 = 4m(m+1) - 4m^2 = 4m = 2n$$

Мы настоятельно рекомендуем читателю вывести подобные отношения для общего случая положительного целого числа **n** в качестве весьма полезного упражнения. На базе данного распределения К. Пирсоном был предложен *наиболее распространенный* на сегодня **χ**-квадрат критерий согласия для проверки статистических гипотез.

***Распределение Стьюдента*** (**t-распределение**) есть распределение случайной величины **T,** определяемой следующей формулой:

$$T = \frac{X}{\sqrt{Y/n}}$$

где **X** и **Y** – независимые случайные величины: **X** имеет *нормальное распределение* с *параметрами* **M[X] = 0** и **D[X] = 1,** а **Y** имеет **χ**-*квадрат распределение* с **n** степенями свободы. Тогда функция плотности **f(t) t**-распределения на открытом интервале (**-∞ < t < +∞**) имеет следующий вид:

$$f(t) = \frac{\Gamma((n+1)/2)}{\sqrt{pn}\,\Gamma(n/2)(1+t^2/n)^{(n+1)/2}}, \qquad \Gamma(n) - \textit{gamma function}$$

**Γ(n)** – гамма-функция. При росте значения **n t-**распределение асимптотически приближается к классическому *нормальному распределению*. Распределение Стьюдента играет весьма важную роль при проверке гипотезы о средней величине нормально распределенной генеральной совокупности при неизвестной дисперсии.

***Распределение Фишера*** (**F-***распределение*) – *распределение случайной* величины **F,** определяемой следующей весьма простой формулой:

$$\mathbf{F = \frac{X/n}{Y/m}}$$

где **X** и **Y** – *случайные величины*, имеющие **χ**-*квадрат распределение* соответственно со *степенями свободы* **n** и **m.** Функция плотности **f(x)** **F**-распределения на открытом интервале (**-∞ < x < +∞**) имеет следующий вид:

$$\mathbf{f(x) = \frac{\Gamma((m+n)/2)\, m^{m/2} n^{n/2} x^{(n-2)/2}}{\Gamma(m/2)\Gamma(n/2)(m+nx)^{(m+n)/2}}}$$

где **Γ(n)** – гамма-функция. Распределение Фишера широко применяется в дисперсионном анализе для оценки достоверности отношения дисперсий и в корреляционном анализе для оценки расхождения между двумя коэффициентами корреляции. Для удобства применения рассмотренных трех критериальных распределений существуют специальные расчетные таблицы, приводящиеся во многих математических справочниках и книгах по теории вероятностей, математической и другим типам статистик. Ниже будут рассмотрены основы статистической проверки гипотез с применением данных критериальных распределений. Наряду с *вышеупомянутыми критериями* имеется целый ряд других статистических *критериев* (Смирнов, Бартлет, Колмогоров-Смирнов, Рени, Хартли, Хотеллинг, Вар ден Ваерден и др.), каждый из которых имеет свои преимущества в соответствующей прикладной области.

Практическое применение математической статистики состоит в, частности, в проверке фактического соответствия реальных результатов экспериментов предполагаемой гипотезе. С этой целью строится процедура проверки гипотезы (*критерий согласия*), позволяющая по результатам наблюдений принимать или отвергать данную гипотезу. Применение критерия согласия сопряжено с ошибками двух типов: отвергнуть справедливую гипотезу (*ошибка первого типа*) и принять ошибочную гипотезу (*ошибка второго типа*). При создании *критериев* согласия стремяться минимизировать ошибки обоих типов. Однако, в большинстве важных на практике ситуаций невозможно построение критериев согласия со сколько *угодно малыми* ошибками *первого* и *второго* типов. Правила проверки гипотез имеют сугубо статистический смысл: при многократном применении *определенного* правила доля числа неверных решений выражается вероятностями ошибок *первого* и *второго* типов. Когда *экспериментальные* данные согласуются с предполагаемой гипотезой, это еще не значит, что невозможно согласование этих же данных с *другой* гипотезой. При применении статистических *критериев* на основании наблюдений невозможïно доказательство той или иной гипотезы. Можно лишь утверждать, что результаты наблюдения не противоречат принятой гипотезе. Таким образом, выводы, принимаемые на основе статистических данных, формулируются следующим образом: результаты статистического наблюдения согласуются с данной гипотезой (*противоречат ей*). Критерии согласия должны удовлетворять ряду требований, подробно рассматриваемых в математической статистике. Для сравнения различных критериев согласия между собой используются меры асимптотической эффективности критериев, которые основаны на изучении скорости сходимости их функций мощности. Из наиболее известных классов критериев согласия можно отметить такие, как: (*последовательные*) критерии отношения правдоподобия, $\chi^2$-критерий, (*порядковые*) *непараметрические* критерии и другие. В частности, $\chi^2$-критерий на практике оказывается достаточно эффективным, лишь когда все ожидаемые частоты удовлетворяют следующему соотношению **n*Pk ≥ 10.** В то время как порядковые непараметрические критерии строятся по статистикам *вариационного ряда* (**ВР**), которые не зависят от конкретных значений членов исследуемого **ВР**.

Более подробно с теорией вероятностей и ее прикладными аспектами можно ознакомиться в книгах [57, 58, 62, 63, 161-163, 178, 181, 183, 296]; при этом, в книге [62] дан весьма интересный

подбор вероятностных задач с подробным разбором их решений. Тогда как книга [188] содержит достаточно большое количество *ссылок* на различные разделы теории вероятностей и статистики. Данная книга представляет существенный интерес какдля исследователя, так и для любого специалиста, использующего в своей профессиональной работе вероятностно-статистические методы. Более того, учитывая вклад, сделанный российскими и советскими школами по теории вероятностей, математической статистике и другим видам статистики, может представить определенный интерес и книга [189].

В настоящее время, пожалуй, нет области знания, в которой не использовались бы методы *стохастики*, которая соединяет элементы теории вероятностей и математической статистики. Применение вероятностно-статистических методов стало традиционным в физике, технике, биологии, экономике, медицине, психологии, социологии, теории обучения и т. д. Теория вероятностей является математической основой статистики , а развитие статистических и кибернетических идей, в свою очередь, способствовало еще большему возрастанию прикладного значения теории вероятностей в качестве теоретической основы многих важных приложений, включая статистику в целом, и специальные статистики в частности.

Область приложений курса *теории вероятностей* и *математической статистики* составляет анализ количественных закономерностей массовых случайных явлений, что обусловило его изучение студентами практически всех специальностей. Изучение теории вероятностей, как правило, является заключительным этапом в освоении блока математических дисциплин и, следовательно, позволяет продемонстрировать студентам возможности применения ранее полученных математических знаний не только внутри данного блока, но и на практике, т.е. реализовать как внутри-, так и межпредметные связи. Профессиональная направленность курса *теории вероятностей* реализуется, прежде всего, через его содержание, а именно через систему прикладных задач, материалом которых служат сведения, полученные студентами из специальных дисциплин. Такое построение курса не только демонстрирует возможности математических методов исследования, но и создает прочную основу для последующего изучения статистических дисциплин.

В любом случае, владение основами теории вероятностей и математической статистики позволяет более сознательно работать в сугубо прикладных и общей статистике, тогда как при отсутствии этих знаний работа в статистических приложениях будет носить, во многом, схоластический характер и творческий аспект в этой связи будет либо весьма слабым, либо и вовсе отсутствовать. Более того, серьезный уровень работ даже в общей и прикладных статистиках настоятельно требуют владения указанными предметами.

# Глава 3.
# Основы статистического наблюдения

***Статистическое наблюдение*** (*или первичный статистический учет*) является специально организованной регистрацией признаков каждой единицы изучаемого объекта и записью их в определенных документах. *Статистическое наблюдение* (*в дальнейшем просто* ***наблюдение***) – *первый* этап статистического *анализа*, направленный на *получение достоверной* статистической информации для *последующей характеристики* изучаемого явления при помощи *обобщающих* статистических показателей, необходимых *оперативному* руководству и *управлению* народным хозяйством, его планированию и прогнозированию. Задача наблюдения состоит в *планомерно* организованном сборе *достоверных* сведений о каждой единице изучаемого *явления* в *условиях* конкретных времени и места.

В процессе наблюдения формируется первичная статистическая информация, подвергаемая затем систематизации, сводке, обработке, анализу и обобщению. От качества наблюдения зависит успех всего *статистического анализа* в целом, поэтому оно должно быть организовано так, чтобы в результате его проведения были получены объективные и точные данные об изучаемом явлении. Перед наблюдением следует весьма четко определить *объект* и *единицы* наблюдения. ***Объектом*** (*совокупность промышленных предприятий, больниц и т.д.*) *наблюдения* называется *совокупность*, являющаяся предметом исследования. *Определить объект наблюдения* – значит указать его *основные отличительные, важнейшие* признаки. Каждый объект *наблюдения* состоит из отдельных единиц. Характеристика объекта может быть получена лишь на основе характеристики его единиц. ***Единицей*** (*завод, вуз, НИИ, научный работник и др.*) *наблюдения* называют составной элемент объекта наблюдения, который является носителем признаков, подлежащих регистрации.

***Единица наблюдения*** – *первичный* элемент объекта статистического наблюдения, являющийся носителем регистрируемых при наблюдении признаков. Для каждого наблюдения должна быть четко определена единица наблюдения и ее отграничивающие признаки, что позволит четко идентифицировать исследуемый объект в целом. ***Единица совокупности*** – носитель признаков, подлежащих регистрации, тогда как *единица наблюдения* – источник сведений, получаемых в процессе наблюдения. В общем случае эти два понятия различны, но иногда совпадают. Например, при переписи населения отдельный человек является и единицей наблюдения, и единицей совокупности *одновременно*. Подобно *объекту*, его *элементы* требуют обоснованного определения.

В зависимости от задачи наблюдения формулируется ***цель*** наблюдения, которая должна быть *конкретной*, *актуальной* по содержанию и направлена на сбор *достоверной* информации, необходимой для анализа, планирования и прогнозирования. Цель наблюдения может быть как *узконаправленой*, так и более широкого назначения. Как правило, *наблюдение* предполагает следующие три последовательных этапа, а именно: (**1**) *разработка программы наблюдения по каждой интересующей нас единице*; (**2**) *определение содержания плана сбора и обработки данных о признаках единицы наблюдения*; (**3**) *контроль результатов наблюдения*. От корректной постановки

цели зависит содержание программы и плана наблюдения изучаемой совокупности, иначе выводы последующего статистического анализа могут оказаться недостоверными.

## 3.1. Программа и план статистического наблюдения

***Программа наблюдения*** представляет собой перечень вопросов, на которые должны быть получены ответы по единицам исследуемого объекта. Содержание программы и количество вопросов зависит от особенностей изучаемого явления и цели наблюдения, направленного на получение предварительно разработанной системы обобщающих показателей. ***Обширная*** программа, содержащая большое число вопросов, позволяет получать более разнообразные сведения и поэтому более углубленно исследовать сущность и особенности изучаемой совокупности единиц. ***Узкая*** программа ориентирована на сбор сведений по ограниченному кругу вопросов, часто имеющих лишь *сугубо* практическое значение. ***Специальная*** программа ориентирована на учет специфических особеностей объекта исследования, например, сбор информации о публикациях по определенному разделу науки с целью проведения его статистического анализа.

***Разработка программы*** – одна из важнейших методологических и прикладных проблем наблюдения, ибо от ее качества зависит ценность собранной статистической информации. Она, как правило, достаточно сложна и требует весьма широкого участия практических и теоретических работников, заинтересованых в результатах такого наблюдения. К программе *наблюдения* и ее *разработке* предъявляется ряд требований, которым она *должна удовлетворять* при любом статистическом анализе. Отметим только основные из них:

- *программа должна содержать существенные признаки, непосредственно характеризующие изучаемый объект*
- *программа должна содержать признаки, позволяющие при необходимости проводить более глубокий анализ объекта*
- *в программу не следует включать второстепенные признаки, затрудняющие дальнейшую обработку данных*
- *при невозможности получения полных и достоверных данных следует ограничить объем собираемых сведений*
- *при возможности и целесообразности в программе следует предусмотреть сбор сведений и за период, не входящий в намеченный; данный подход во многих случаях впоследствии поможет облегчить контроль исходных статистических данных*
- *в программу, по-возможности, следует включать вопросы контрольного характера, служащие целям проверки и контроля собранной статистической информации*
- *программа не должна содержать вопросов, на которые не могут быть получены объективные ответы, и нечетко либо двусмысленно поставленных вопросов*
- *целесообразно так формулировать вопросы, чтобы получаемые ответы за исследуемый период были сравниваемы с данными статистических наблюдений предыдущих периодов*
- *программа должна содержать адресную часть, идентифицирующую местоположение источника статистической информации*.

Программа наблюдения оформляется в виде специального статистического бланка или формуляра. Бланк имеет две формы: *индивидуальную* и *списочную*. В первом случае на каждую единицу совокупности выделяется отдельный бланк, во втором – для каждой единицы в бланке отводится отдельная строка или графа. Применение современной **ВТ** и средств связи для обработки статистической информации выдвигает ряд серьезных требований к форме статистических бланков и формуляров, т.к. они должны являться и носителями вводимой в ЭВМ первичной информации. В качестве их могут выступать магнитные ленты, диски,

дискеты и другие магнитные носители информации. Для обеспечения единообразия в толковании программы наблюдения разрабатываются соответствующие инструкции по объяснению и взаимному контролю ее вопросов. Инструкция оформляется либо отдельной брошурой, либо на самих бланках. Вся документация по наблюдению (*бланки*, *формуляры*, *инструкции и др*.) образуют *инструментарий* статистического наблюдения.

В целях успешного проведения наблюдения необходимо разработать его методологическо-организационный план, в котором отражается решение важнейших вопросов организации и проведения наблюдения с указанием конкретных сроков и места проведения намеченных мероприятий. В плане наблюдения различают *программно-методологические* и *организационные* вопросы, а именно:

1. ***Программно-методологическая часть:***
   - определение и разграничение цели и объекта наблюдения;
   - определение единиц совокупности и наблюдения;
   - разработка программы наблюдения;
   - определение типа и вида наблюдения;
   - разработка бланков наблюдения и инструкций по их заполнению;
   - разработка методов контроля результатов наблюдения;
2. ***Организационная часть:***
   - определение состава наблюдателей (регистраторов);
   - обучение и инструктаж наблюдателей;
   - размножение и рассылка документов наблюдения;
   - порядок проведения наблюдения, приема и сдачи его материалов;
   - порядок получения и представления предварительных и окончательных итогов;
   - проведение (при необходимости) пробного наблюдения.

Для корректной характеристики изучаемого объекта весьма важное значение имеет выбор времени и места наблюдения. Временные границы устанавливают в зависимости от динамических особенностей самого объекта и цели исследования. Период, в течение которого производится сбор данных, называется ***периодом наблюдения***. Он отличается от *периода регистрации*, за который эти данные собираются. Сам момент регистрации (*сбора*) данных называется *критическим*. Например, при *Всесоюзной* переписи населения *критическим* моментом была выбрана полночь 12 января 1989 г., а период наблюдения составлял **8** дней (13 - 20 января 1989). Временные границы устанавливаются только для единовременных наблюдений. Критический момент требуется в случае интенсивной динамики исследуемого объекта и выбор его следует приурочить к *минимуму* его *динамики*. Вопрос о *месте проведения* наблюдения (***территория наблюдения***) приобретает особую актуальность при исследовании *перемещающихся* объектов. Здесь также следует определить *наиболее вероятное местоположение* объекта. Например, перепись населения охватывает всю страну, а единица наблюдения – человек регистрируется по месту его постоянного жительства.

## 3.2. Основные формы, виды и способы статистического наблюдения

Все многообразие статистических форм наблюдения по его цели можно разбить на две большие группы: ***первично*** и ***вторично*** статистические. К *первой* группе относятся наблюдения с целью получения только статистической информации, например, перепись населения. Ко *второй* – наблюдения, основная цель которых не статистика, а некоторая иная.

Так, например, *бухгалтерский учет* служит для контроля финансово-хозяйственной деятельности, оформляется отчетами, часть информации из которых используется статистикой. Особую же роль в деле получения бухгалтерской отчетности играют *централизованные бухгалтерии* (**ЦБ**), на которые возложена организация учетно-отчетных работ однородных, небольших по размерам предприятий, организаций или учреждений. Наиболее развитой системой **ЦБ** характеризуются сфера бытового обслуживания, образование, здравоохранение и др. Ко второй группе относится большинство специальных наблюдений в различных областях (*судебное дело*, *налоговая служба*, *здравоохранение*, *наука*, *образование*, *социология и др.*).

По временному фактору наблюдения деляться на ***текущие***, ***повторяемые*** и ***единовременные***. *Текущее* наблюдение производится систематически, непрерывно или через небольшие интервалы времени. Например, регистрация актов гражданского состояния в ЗАГСах, статистика товарооборота в торговле, на транспорте и др. К этой же группе относится статистическая отчетность (*месячная*, *квартальная*) предприятий, организаций и учреждений, оформленная в виде обязательных отчетов об их деятельности. *Отчетность* является важнейшей формой наблюдения и имеет большое значение, являясь одним из основных источников статистической информации о народном хозяйстве и развитии страны в целом. *Повторяемое* наблюдение проводится или через *равные* интервалы времени, или *нерегулярно* по мере надобности. Например, ежегодная отчетность вузов о приеме-выпуске студентов, ежегодные отчеты о ходе уборки урожая. *Нерегулярным* наблюдением, например, является учет нанесенного ущерба в результате некоторого стихийного бедствия. *Единовременное* наблюдение проводится или один раз, или повторяется через неопределенные промежутки времени, например, переписи населения, основных фондов страны, жилого фонда и др. Преимуществом текущих и повторяемых наблюдений (*в отличие от единовременных*) является возможность исследовать динамику объектов.

По полноте охвата единиц совокупности наблюдения деляться на ***сплошные*** и ***несплошные***. При *сплошном наблюдении* регистрации подвергаются *все единицы* исследуемой совокупности, например, в случае переписи населения или бухгалтерского учета. В частности, недавняя перепись населения и жилищного фонда Эстонии (31 *марта* – 9 *апреля* 2000) охватили всех жителей страны и участие в этом мероприятии было фактически *принудительным*. *Анкетный* опрос включил 43 вопроса (3 *факультативных*, 9 *об условиях жизни интервьюируемого и* 31 *о других аспектах его состояния*).

Во втором случае регистрации подвергается только часть единиц совокупности, но выводы обобщаются на всю совокупность. При этом, заранее при планировании наблюдения учитываются такие моменты как: (**1**) *сам факт несплошной регистрации*; (**2**) *какая часть совокупности выбирается*; (**3**) *способ отбора исследуемых единиц*. *Несплошное* наблюдение по сравнению со *сплошным* требует гораздо меньше времени, сил и средств, позволяет использовать более подробную программу и более совершенный способ учета, быстрее подводить итоги и повышает оперативность статистической информации. В некоторых случаях *несплошное* наблюдение является единственно возможным (*контроль качества продукции*, *анализ спроса и предложения в торговле и др.*). В статистической практике используется несколько видов несплошного наблюдения, из которых основными являются: *монографическое*, *сравнительно-монографическое*, *метод основного массива*, *выборочное* и *анкетное*.

***Монографическое*** наблюдение состоит в обследовании одной, *типичной для всей совокупности единицы*, с обобщением полученных результатов на всю совокупность. Такой вид наблюдения имеет наибольшее применение в ботанике и зоологии. Он также широко используется для распространения передового опыта и установления причин недостатков

путем обследования соответственно передовых и отстающих предприятий, а также выявления имеющихся или намечающихся тенденций. При статистическом анализе экономических явлений наиболее эффективным является ***сравнительно-монографическое*** *наблюдение*, состоящее в *обследовании* в каком-то смысле крайних единиц совокупности и сравнении полученных результатов. Так, например, обследованию подвергаются самое передовое и самое отсталое предприятие отрасли.

При методе ***основного массива*** обследованию подвергаются наиболее *существенные* единицы изучаемой совокупности. Например, наблюдение за *ценами* на городских рынках проводится в 264 наиболее крупных городах страны, составляющих менее **5%** всех городов, но население которых составляет более половины всего городского населения страны. При ***выборочном*** наблюдении характеристика всей совокупности производится по некоторой (*сравнительно небольшой*) ее части, отобранной случайным образом. При этом, вся совокупность называется ***генеральной,*** а ее обследуемая часть **–** ***выборочной совокупностью***.

***Выборочное*** наблюдение – самый широко распространенный вид несплошных наблюдений в статистике. Во многих случаях мы можем заменить *сплошное* наблюдение соответствующим *несплошным* наблюдением при условии правильной организации и реализации наблюдения. Вычисление типичного размера и вероятности, с которой полученные результаты являются корректными относительно генеральной совокупности, производятся методами теории вероятностей. Основные принципы метода реализации *несплошного* наблюдения (*выборочного наблюдения*) будут рассматриваться ниже более подробно.

Суть ***анкетного*** *наблюдения* состоит в рассылке *специальных анкет* источникам статистической информации с последующим их получением заполненными. На практике, процент возврата анкет не слишком велик и этот метод, как правило, используется в социологических опросах, различных опросах читателей газет и журналов и др. К анкетному методу непосредственно примыкает и метод ***интервью,*** когда опрос ведется путем личного общения по заранее разработанному опроснику. По способу учета регистрируемых фактов наблюдения делятся на *непосредственные*, *документальные* и различного рода *опросы*.

При ***непосредственном*** наблюдении статистическая информация получается путем личной регистрации единиц совокупности, например, регистрация температуры внешней среды, потока пассажиров в метро, инвентаризации разного рода и др. Получаемые в результате непосредственного наблюдения материалы достоверны, но их получение весьма трудоемко и дорогостояще. ***Документальное*** наблюдение базируется на использовании в качестве *источника* статистической *информации* различных документов *первичного учета* предприятий, учреждений и организаций (*регистры бухгалтерского учета, отчеты и др.*). Наибольшая точность и достоверность статистической информации достигается при *непосредственном* и *документальном* наблюдениях. Достоверность статистической информации (*порой весьма существенно*) меньше именно при *опросном* наблюдении: *устном*, *саморегистрации* и *корреспондентском*.

При ***устном*** (*экспедиционном*) способе регистратор производит опрос и со слов обследуемого заполняет опросный бланк, что гарантирует единообразное понимание вопросов. Несмотря на большие затраты сил и средств к нему приходится прибегать, поскольку в ряде случаев он является *единственно* возможным. При способе *саморегистрации* также используется институт регистраторов, но бланки заполняются самими опрашиваемыми. Обязанность регистраторов состоит в раздаче бланков опрашиваемым, их инструктаже, сборе и проверке заполненных бланков. Данный способ используется при бюджетных обследованиях семей, проведении некоторых видов *переписей* и др. ***Корреспондентский*** *способ* заключается в том, что *специально* разработанные бланки и инструкции к ним рассылаются отдельным организациям или

специально подобранным лицам, давшим согласие периодически заполнять их и присылать обратно в *установленные* сроки. Преимуществом данного способа является его относительная дешевизна, однако качество его сведений всецело зависит от *квалификации* и *добросовестности* корреспондента. Примером такого способа статистического наблюдения является созданная ВНИИКСом широко разветвленная сеть корреспондентов на территории бывшего Союза. В связи с созданием информационно-компьютерных сетей во многом меняются методы и организация сбора и доставки в статистические органы результатов наблюдения, особенно в области различного рода отчетности. В первую очередь, это касается методов доставки информации, ее *контроля* и *первичного* статистического анализа. *Современные* компьютерные технологии позволяют делать первичную обработку статистической информации в местах ее зарождения. Таким образом, более низкие уровни статистического анализа характеризуются увеличивающейся тенденцией к децентрализации.

## 3.3. Вопросы точности статистического наблюдения

Несмотря на тщательность подготовки при проведении наблюдения, могут иметь место ошибки, приводящие к снижению его достоверности. ***Ошибка наблюдения*** – расхождение действительных значений признаков единиц совокупности с их значениями, полученными в результате наблюдения. Чем больше значение ошибки, тем меньше точность наблюдения. Можно выделить следующие основные виды ошибок статистического наблюдения.

***Ошибки репрезентативности*** свойственны только выборочному наблюдению; их нельзя избежать, но можно достаточно точно определять их значения. Причиной их появления является то, что при формировании выборочной совокупности, практически, невозможно получить структуру, идентичную структуре генеральной совокупности. ***Случайные*** ошибки репрезентативности возникают в силу несплошного характера наблюдения (*неполно воспроизводящего генеральную совокупность*) и могут быть удовлетворительно оценены. ***Систематические*** ошибки возникают по причине нарушения принципов случайного отбора единиц при формировании выборочной совокупности и их размеры обычно не поддаются удовлетворительной количественной оценке.

*Ошибки* ***регистрации*** возникают из-за *неверного* установления фактов в процессе наблюдения, ошибочной их записи или того и другого *одновременно*. ***Преднамеренные*** ошибки возникают в результате умышленного искажения фактов (*искажение отчетности*, *занижение показателей*, *приписки и др.*). Выявление ошибок такого рода влечет за собой административную или уголовную ответственность виновных. ***Непреднамеренные*** *ошибки* совершаются *неумышленно* и возникают по различным причинам: *неправильное понимание вопросов программы наблюдения*, *невнимательность участников наблюдения* и др. *Случайные непреднамеренные ошибки регистрации* данных могут, полагают, с одинаковой вероятностью искажать результаты наблюдения в противоположные стороны. Они в массе единиц не оказывают существенного влияния на конечные результаты наблюдения, ибо в процессе статистической сводки они, как правило, взаимопогашаются. Систематические ошибки искажают сведения по отдельным единицам совокупности в одном направлении, например: ошибки округления возраста; ошибки, возникающие из-за неисправности измерительных приборов и др. Во избежание подобных ошибок следует обратить, в первую очередь, особое внимание на разработку программы наблюдения и инструкции к ней, подбор и подготовку кадров регистраторов, задействовать надежные методы верификации результатов статистического наблюдения. Наша практика показывает, что хорошая проработка перечисленных моментов весьма существенно сказывается на достоверности получаемой в процессе статистического наблюдения информации, сводя к минимуму перечисленные выше типы ошибок.

## 3.4. Контроль результатов статистического наблюдения

Во избежание ошибок наблюдения, выявления, исправления и уменьшения их размеров необходимо в процессе подготовки и проведения наблюдения предусмотреть и осуществить ряд мероприятий. К такого рода мероприятиям, например, относятся: (**1**) четкая разработка цели, задачи и программы наблюдения; (**2**) детальная инструкция по *программе* наблюдения, содержащая методику вычисления изучаемых показателей; (**3**) разумный выбор контрольной даты наблюдения, его сроков и места; (**4**) правильный подбор кадров регистраторов и их обучение. В целом ряде случаев целесообразно вводить особые мероприятия, позволяющие проводить проверку наблюдения в ходе его выполнения. Так, например, при переписи населения 1989 г. использовалась выдача справок о прохождении переписи с заполнением контрольных бланков. Так, по результатам проверки контрольных бланков в переписные листы дополнительно было внесено 453000 человек. Большое значение для получения *достоверной* статистической информации имеет действенный контроль за ходом *наблюдения*, систематические проверки состояния первичного учета и отчетности на предприятиях, в учреждениях и организациях, а также серьезные меры борьбы с *сознательными* искажениями отчетности. Это особенно актуально в настоящий переходный период, когда работа многих контролирующих органов испытывает весьма серьезные затруднения, а число источников статистической информации различной принадлежности многократно возросло. Общая социально-экономическая ситуация (*включая ее серьезные криминальные аспекты*) также играет весьма отрицательную роль в деле приобретения достоверной статистической информации и принятия решения на ее основе.

Более того, процесс наблюдения должен завершаться ***контрольными мероприятиями,*** позволяющими проверять и повышать точность полученных данных. В качестве их можно использовать, например, частичные или сплошные контрольные проверки результатов наблюдения, осуществляемые сразу после его окончания. Однако такой подход из-за его трудоемкости и дополнительных затрат используется по мере его целесообразности. По окончании статистического наблюдения собраные данные должны быть весьма тщательно проверены. Проверка производится по следующим основным направлениям: (**1**) полнота охвата объекта наблюдения, (**2**) качество *заполнения бланков* и других *документов* наблюдения, а также (**3**) само существо статистических данных. В последнем случае различают два вида контроля: *логический* и *арифметический*.

Контроль ***полноты охвата*** возможен и необходим при отчетном способе наблюдения, ибо в противном случае сведенный материал может оказаться неполным, а данные об изучаемой совокупности неточными. При других способах наблюдения контроль полноты и меры его обеспечения производятся, исходя из конкретных *условий* и *специфики* объекта исследования.

***Логический*** контроль ведется с целью проверки корректности самого содержания данных, собранных по каждой единице совокупности. Способы такого вида контроля достаточно разнообразны, а именно:

- *сопоставление ответов на различные вопросы в рамках одного опросного бланка*
- *сопоставление полученных данных, относящихся к отчетному периоду, с аналогичными данными предыдущих периодов и/или плановыми данными отчетного периода*
- *соответствие полученных данных их внутренней сущности*
- *сравнение фактических данных наблюдения с нормативными*
- *сравнение данных наблюдения с результатами специальных наблюдений выборочного типа.*

Сама специфика наблюдения, как правило, определяет целый ряд возможностей проведения тех или иных видов *логического* контроля статистической информации.

***Арифметический контроль*** состоит в проверке различных арифметических расчетов, результаты которых приводятся в бланке наблюдения (*расчет средних*, *контрольные суммы*, *вычисления процентов или взаимосвязей между показателями и др.*). Для успешного проведения *логического* и *арифметического* контроля при разработке программы и плана наблюдения следует предусмотреть ряд *дополнительных* контрольных процедур (*использовать контрольное суммирование*, *допустимые диапазоны значений данных*, *взаимосвязи между показателями и др.*), которые в момент верификации (*а в ряде случаев и в момент проведения самого наблюдения*) статистических данных позволят достаточно надежно выявлять или исправлять ошибочные данные. Задача логического и арифметического контроля существенно облегчается при использовании **ВТ** и средств связи.

С ***достоверностью*** данных наблюдения тесно связан вопрос о целесообразной мере их точности. Его суть состоит в том, чтобы определить оптимальное число значащих цифровых разрядов в *количественном* показателе, которое обеспечивало бы необходимую достоверность данных (*в смысле накопления ошибок округлений*) и в то же время соответствовало бы реальной достоверности получаемых данных. Различают точность *формальную*, *реальную* и *фиктивную*. ***Формальная*** точность определяется значением единицы последнего учитываемого разряда в количественном значении показателя. Так, например, в показателе численности населения СССР на 15.01.70 – 241748 тыс. человек, мерой формальной точности является тысяча человек. ***Реальная*** же точность измеряется значением единицы последнего разряда, достоверность которого гарантируется. Реальная точность не может превышать формальную, в противном случае она является ***фиктивной***. Во избежание этого следует руководствоваться правилом: точность результатов вычислений не должна превышать точности исходных данных; если исходные данные имеют разную точность, тогда точность результата устанавливается по наименьшей точности исходных данных. Различают также *достижимую* и *необходимую* точность. ***Достижимой*** называется максимально допустимая в процессе сбора, контроля и обработки данных точность. ***Необходимая*** точность определяется назначением показателя и его последующим использованием. Из возможных соотношений для *реальной* (**Р**), *необходимой* (**Н**) и *достижимой* (**Д**) точности наилучшим будет **Н=Р ≤ Д.** В случае **Н < Р ≤ Д** имеет место *излишняя* точность, а при **Р < Н ≤ Д** и **Р ≤ Д < Н** следует добиваться повышения *реальной* точности до *необходимой* и/или *достижимой*.

Детальное обсуждение понятия *достоверности* и *ошибок* статистических показателей, а также новая классификация ошибок, основывающаяся на анализе содержания и формы последних, даны в [112]. В кданной ниге рассматриваются интересные вопросы, относящиеся к проблеме контроля достоверности показателей, способам обнаружения и мерам по предотвращению и/или уменьшению возможных ошибок. В условиях массового применения **ВТ,** особенно класса **ПК,** появляется достаточно реальная возможность проведения все более сложных типов *логического* и *арифметического* контролей результатов статистического наблюдения как в процессе проведения уже *первичного* статистического анализа, так и в процессе проведения статистического анализа в целом. Данная работа на сегодня вполне доступна на рабочем месте статистика любого уровня.

## 3.5. Специальные вопросы отчетности и переписи

***Статистическая отчетность*** базируется, в основном, на *первичном* и *бухгалтерском* учетах. Данные виды учета являются важнейшим источником информации, необходимой для заполнения форм отчетности. ***Первичный учет*** – это регистрация различных предметов, событий (*например*, *актов гражданского состояния*), проводимая систематически по мере их возникновения. Она производится, как правило, на особом документе – *первичном учетном*

*формуляре*. В функции первичного учета входят только операции наблюдения (*регистрации*) данных и самый элементарный подсчет итогов. *Первичный учет* – неотъемлемая функция всех предприятий, организаций и учреждений; его программа и принципы определяются потребностями как самого объекта учета, так и вышестоящих органов.

На предприятиях организован учет трех типов: *оперативно-технический*, *бухгалтерский* и *статистический*. ***Оперативно-технический*** учет ведется на основе текущих первичных документов, содержание которых позволяет получать сведения о ходе производственного процесса; он направлен на получение материалов для характеристики единичных явлений (*время простоя рабочих и механизмов, количество обслуженных клиентов и др.*). Данные этого типа учета служат основой для более глубокого статистического анализа объекта. ***Бухгалтерский*** же учет отражает в первичных документах финансово-хозяйственную деятельность объекта. ***Статистический*** учет имеет своим объектом массовые общественные явления, сведения о которых необходимы для получения обобщающих статистических показателей.

Различают отчетность *общую* и *специальную*. ***Общая*** отчетность содержит однотипные данные для определенной отрасли или всего народного хозяйства в целом, тогда как в ***специальной*** содержатся специфические показатели отдельных отраслей. Наряду с ними различают также *общегосударственую* и *ведомственную* отчетности. Первая собирается и подвергается сводке, обработке и анализу в органах государственной статистики для нужд всего народного хозяйства, тогда как вторая собирается министерствами, департаментами и ведомствами, как правило, для своих нужд. По времени представления отчетность делится на *годовую* (*годовой отчет*) и *текущую* (*ежедневную*, *пятидневную*, *декадную*, *месячную* и *квартальную*). Всем государственным и другим органам запрещено требовать, а руководителям предприятий, учреждений и другим представлять отчетность по неутвержденным в *установленном* порядке формам. Борьба с незаконной, неутвержденной отчетностью, а также с ее непредставлением или представлением в искаженном виде является обязанностью статистических органов различного уровня. Материалы специально организованных наблюдений часто дополняют содержание форм отчетности. Статистическая отчетность и специальные наблюдения – основные источники информации о деятельности предприятий, организаций и отраслей хозяйства страны. В СССР вопросами централизации статистической отчетности ведала система ЦСУ (*Центральное Статистическое Управление*), что позволяло успешно повышать эффективность статистической работы в масштабе всей страны, устранять дублирование сбора и обработки отчетных данных ведомствами. В Эстонии эти функции в настоящее время выполняет *Департамент Статистики*.

***Переписи*** – наиболее крупные и сложные специально организованные наблюдения, из которых исторически первыми являются переписи населения. Помимо них проводятся переписи: промышленности, оборудования, земельных угодий, жилого и нежилого фондов и др. Для проведения таких переписей статистикой выработаны особые методология и практика их проведения, обеспечивающие получение сопоставимых данных. В период Великой Отечественной войны статистика выработала особую систему ***срочных*** переписей для срочного получения необходимых статистических данных (*остатки различных материалов, энерго- и электрооборудования, автомобильного и ж/д транспорта и др.*). Программы таких переписей, как правило, содержат небольшое число вопросов. Организационно *срочные* переписи отличаются от обычных переписей тем, что устраняются посредники между предприятиями и государственной статистикой. Местные статистические органы выполняют лишь функции контроля, а получение указаний и инструкций, а также сбор данных производятся обычно по средствам связи. Сводка всех материалов производится централизованно в департаменте статистики. После распада бывшего СССР статистические

наблюдения типа переписей должны быть весьма актуальны в государствах, появившихся на его обломках, хотя бы с целью инвентаризации всех оставшихся ресурсов как людских, так и материальных, ибо без этого достаточно сложно представить полную картину стартового состояния новой страны. Данная задача должна возлагаться на статистические органы различного уровня и решаться в самые сжатые сроки, ибо налицо не только разбазаривание бывшего государственного имущества (*и*, *в первую очередь*, *принадлежащего военному ведомству*), но и прямые факты его незаконного отчуждения в пользу частных лиц и отдельных фирм.

## 3.6. Данные, использованные для иллюстрации рассматриваемого материала

Для иллюстрации основных положений и процедур статистического анализа данных используются те или иные статистические данные, представляющие абстрактные или конкретные совокупности либо выборки из них. Как правило, в качестве этих данных авторы используют статистические выботки из собственной профессиональной деятельности. В нашем случае в качестве исходных данных для иллюстративных примеров рассматриваемого материала выбраны статистические данные, отражающие *творческую* активность *Таллиннской Творческой Группы* (**ТТГ**). В качестве характеристик активности выбраны такие показатели как *количество* и *тип* публикаций наряду с их *объемом* и *местом издания*, *цитирование* публикаций в отечественных и иностранных источниках и т.д. В более широком смысле под понятиями "*отечественные*" и "*иностранные*" в дальнейшем мы будем понимать все, что касается России совместно с прежним СССР и другого цивилизованного мира соответственно.

Более того, в качестве материала для анализируемой совокупности были выбраны научные публикации **ТТГ** в течение 1970–2000 г.г. [23,29,127,190-192]. Данная совокупность собиралось членами **ТТГ** в течение всего времени ее творческой активности. Для этих целей нами использовались различные источники: множество (**1**) периодической и (**2**) непериодической литературы по *однородным структурам* (*Cellular Automata*) и их многочисленным *приложениям* наряду с реферативными журналами, включая *Science Citation Index* (**SCI**) и другие подобные источники цитируемых публикаций. Данная работа наряду с большой трудоемкостью, между тем, позволила существенно адаптировать планирование активности **ТТГ** в данном направлении. Действительно, данная работа потребовала информационного архивирования по публикациям **ТТГ** во всех доступных и отечественных, и иностранных информационных источниках и, прежде всего, в реферативных журналах *различного направления*. Информация, собранная в процессе данной работы, позволила держать руку на пульсе развития данного направления современной кибернетики и вовремя реагировать на изменения его основных тенденций, методов исследования наряду с приложениями в самых различных областях естествознания. Наконец, данная регистрация и статистическая работа позволили собрать уникальную литературу и материалы для многих обзоров по теории *Однородных Структур* (*Cellular Automata*) и ее многочисленным приложениям. Более того, анализ информации такого рода позволяет получать и определенные *прогностические* выводы относительно дальнейшего развития данного направления современной кибернетики. Естественно, в настоящее время данные показатели существенно изменились в сторону увеличения. Однако, по ряду различных причин, статистика по творческой активности **ТТГ** в период 2000 – 2004 г.г. не собиралась. Между тем, данное обстоятельство совершенно не умаляет иллюстративные возможности представленных первичных статистических данных.

При выборе вышеуказанных иллюстративных данных мы исходили из того факта, что в целом ряде случаев предлагаемые иллюстративные данные по статистике носят достаточно специфический характер и не настолько интересны для широких кругов читателей. Кроме того, статистический анализ этих данных с последовательной интерпретацией полученных результатов в ряде случаев предполагает наличие определенного знания читателя в области,

из которой эти данные были взяты. В то время как данные, характеризующие творческую активность, являются достаточно прозрачными и понятными, практически, для любого читателя и не требуют какого-либо дополнительного знания. Более того, исследование науки в целом в рамках науки наук (*sociology of science*) проводится с позиций рассмотрения ее как объекта производства, имеющего специфическое трудовое управление и служащих, которые производят специфическое производство, в значительной степени характеризуемое (*и непосредственно, и косвенно*) публикациями различного типа, уровня и назначения. Более того, в этом отношении "*наука наук*" (*исследование самого предмета исследований*) достаточно сильно примыкает к социальным наукам типа экономики, социологии и т.д. По этой причине, материал (*включая иллюстративные примеры, предлагаемые в настоящей книге*) будет не безинтересен для круга читателей, на который данная книга и ориентируется. Кроме того, статистический анализ данных такого типа наряду с сугубо иллюстративными целями преследует также ряд более практических целей, которые имеют определенный интерес для такой быстро развивающейся в настоящее время дисциплины как *социология науки*.

На основе статистического анализа творческой активности даже одной научной организации (*типа, например*, **ТТГ**) могут быть получены достаточно интересные выводы. Результаты, полученные в этом направлении, могут быть достаточно полезны при изучении такого важного объекта как наука, которая теперь достаточно активно исследуется как на ее настоящем уровне развития, так и относительно ее перспектив на самую близкую и более долгосрочную перспективу [208-212]. Статистический анализ такого типа данных имеет определенное значение как для разработки, так и для планирования научных и числовых моделей прогноза в естественно научных дисциплинах.

В научном исследовании (*как процессе обработки информации*) возможно выделить довольно определенные параметры, отражаемые временным рядом. Если оценки их значений задаются некоторыми функциями распределения вероятностей, то возможно строить поведенческую модель творческого процесса на некотором временном интервале в будущем. Однако, вся сложность состоит в определении в научном исследовании начальных параметров, особенно структурных. В настоящее время, есть более или менее развитые методы функций автокорреляции и экспоненциального сглаживания для предсказания экономических процессов; однако, и они требуют дальнейших усовершенствований и развития. Относительно научной творческой активности мы имеем только некоторые критерии определения старения фактического научного знания, т.е. можем отмечать только возникающие переоценки значения существующей информации. Это позволяет определять специфические "*точки роста*" научного знания, в котором развитие знания принимает кризисный характер и в котором, прежде всего, необходимо ожидать научные открытия. Можно отметить следующие методы определения научного значения информации в условиях проблематичной ситуации, а именно: степень силы прогноза, статистические показатели *количества публикаций* и их *цитирования*, и т.д. Примеры, рассматриваемые в настоящей книге относительно развития такой области современной кибернетики как *Однородные Структуры* (*Cellular Automata в английской терминологии*), позволяют в *определенной* мере прояснить некоторые из этих проблем.

Один из главных показателей *творческой активности* научной организации – цитирование ее публикаций. Контроль цитирования научных публикаций позволяет наблюдать динамику развития той или иной идеи и ее проникновения в другие дисциплины. В случае получения кривых цитирования в течение ряда лет появляется хорошая возможность определения динамики роста активности научной организации или отдельных исследователей. На основе такой информации было бы возможно оценивать динамику цитирования отдельных работ и результатов научной активности в целом наряду с продолжительностью периода скрытого

развития (*лага*), когда работа пока еще не цитировалась, порядок роста при движении к максимуму цитирования, степень максимальной пологости и т.д. Сравнение уровней цитирования научных публикаций в *отечественных* и *иностранных* источниках представляет особенно важный аспект, который непосредственно связан с проблемой лингвистической изолированности науки. Так, например, в целом ряде стран в некоторых областях науки цитируются главным образом работы, изданные на языке той же самой страны. В то время как относительно других областей, картина может быть совершенно обратной.

Данные, используемые для иллюстративных примеров, относятся к научной активности только одной научной организации (*а именно*, **ТТГ**); однако, даже их статистический анализ позволяет получать ряд интересных результатов, непосредственно связанных с обсуждением вопросов в вышеупомянутых разделах современной социологии науки. Наряду с этим, эти данные представляют довольно прозрачные совокупности для иллюстрации на их основе основных приемов, методов и процедур общей теории статистики для студентов социально-экономических профессий.

Между тем, вместо предложенных статистических данных, используемых для иллюстрации рассматриваемых в нашей книге статистических приемов и статистического анализа данных в целом читатель может аналогичным образом применять и свои данные, ориентированные на свои профессиональные приложения. В значительной мере это позволит более осознанно усваивать предложенный материал с ориентацией на вполне конкретные приложения.

# Глава 4.
# Сводка, группировка и представление статистических данных

***Сводка*** и ***группировка*** – *второй* этап статистического анализа. Если от *первого* этапа анализа – статистического наблюдения – зависит полнота, качество, а также достоверность собранной информации, то от *второго* – эффективность использования данных для решения задач анализа. Под *сводкой* понимается обработка материалов наблюдения для получения *итоговых* или определенным образом *упорядоченных числовых характеристик* изучаемой совокупности. На этой стадии анализа совершается переход от *характеристик единиц* к характеристике их *совокупности*.

При *сводке* основное внимание уделяется именно группировке результатов наблюдения. Под *группировкой* понимается разделение единиц совокупности на однородные группы по существенным варьирующим признакам, называемым *группировочными*. Результаты сводки и группировки данных, как правило, оформляются в табличном и/или графическом виде, причем *первый* вид представления данных наиболее рационален и удобен для дальнейшей обработки, тогда как *второй* дает максимум наглядности. Настоящая глава рассматривает все три указанные компоненты *второго* этапа статистического анализа данных.

## 4.1. Задачи сводки данных и ее содержание

В узком смысле *сводка* представляет собой методику *подведения итогов* в группах и подгруппах совокупности и оформление их в табличном виде. Тогда как в широком смысле ***сводка*** охватывает группировку результатов наблюдения, составление системы показателей для характеристики типичных групп и подгрупп изучаемой совокупности, подсчет числа единиц и итогов в группах и подгруппах, а также оформление результатов в *табличном* виде.

*Сводка* производится уже на основе детального предварительного анализа и по заранее составленной программе. Программа в первую очередь определяет *подлежащее* и *сказуемое*. ***Подлежащее*** сводки составляют группы, а ***сказуемое*** – показатели, характеризующие каждую группу и совокупность в целом. При этом группы совокупности могут быть получены по многим признакам и охарактеризованы многими *показателями*. Можно выделить следующие основные виды сводки по: (**1**) месту проведения (*централизованная и децентрализованная*), (**2**) группировке данных (*простая и сложная*) и (**3**) методу выполнения (*ручная и механизированная*).

При ***централизованной*** сводке весь материал наблюдения обрабатывается в центральном статистическом органе, тогда как при ***децентрализованной*** сводке материал наблюдения проходит поэтапную обработку. Окончательные итоги получаются при сводке результатов разных этапов (*вторичные сводки*). *Децентрализованная* сводка имеет целый ряд преимуществ перед *централизованной*, ибо производится в районе получения *первичных* данных. Примером *децентрализованной сводки* может служить обработка статистической отчетности, а примером

*централизованной* – обработка результатов переписей населения. Понятия ***простой*** и ***сложной*** сводки соответствуют ранее упомянутым понятиям сводки соответственно в *узком* и *широком* смыслах. Понятие ***ручной*** сводки особых пояснений не требует, а ***механизированная*** в настоящее время предполагает использование современных средств **ВТ** (*и, в первую очередь, персональных компьютеров*) и связи, и является основной. *Механизированная* сводка предъявляет ряд требований к упорядочению первичной статистической документации, созданию различных классификаторов и кодификаторов. Основными составными элементами сводки являются следующие:

- *программа, определяющая группировки, применяемые в разработке (**подлежащее**), и система показателей, характеризующих совокупность в целом и ее отдельные группы (**сказуемое**)*
- *подсчет групповых и общих итогов*
- *оформление конечных результатов сводки в статистических таблицах и/или графиках.*

Основное содержание программы сводки составляет система макетов *разработочных таблиц*. Для *успешного* осуществления сводки составляется *план* ее проведения, отражающий вопросы организации сводки, табличного оформления ее результатов, публикации соответствующих статистических сборников и других статистических материалов, представляющих интерес для различных организаций и ведомств.

## 4.2. Основы метода группировок статистических данных

***Группировка*** – *основа методологии* обработки статистических данных. Можно констатировать, что в арсенале статистики *группировка* является одним из наиболее эффективных и мощных средств. Результаты группировки могут быть выражены *групповыми* и *комбинационными* таблицами. Табличное представление сгруппированных данных позволяет изображать сложную картину общественных явлений в целом. Несмотря на кажущуюся легкость проведения группировок и технический характер дела, эта операция является достаточно сложным этапом стаистическогоанализа. Можно выделить следующие основные типы задач, решаемых методом группировок:

- *выделение социально-экономических типов*
- *изучение структуры явления и структурных сдвигов, происходящих в нем*
- *выявление связи и зависимости между явлениями.*

***Группировки*** в статистике решают многие важные задачи, но все они в конечном счете преследуют одну основную цель – упорядочить первичные статистические данные, чтобы подвергнуть их дальнейшему анализу посредством различных статистических методов: средних, относительных величин, индексов, регрессионного и дисперсионного анализов и др. Можно выделить три основных принципа *группировок*, а именно по: (**1**) *целям исследования*, (**2**) *числу* и (**3**) *характеру группировочных признаков*.

В первом случае группировки делятся на три основных типа: *типологические*, *структурные* и *аналитические*. ***Типологическая*** *группировка* делит всю совокупность на *качественно однородные* группы социально-экономического характера. Техника распределения единиц на *типические* группы в целом ряде случаев достаточно сложна и определяется признаком, который следует положить в основание группировки. *Типологические* группировки можно применять там, где нужно характеризовать качественные особенности отдельных групп.

В качестве примера *типологической* группировки приведем распределение публикаций **ТТГ** [23] по типу и месту издания (табл. 1). При этом в качестве материала наблюдения берется *совокупность* научных публикаций **ТТГ** за 1970-1999 годы [23,29,127,190-192]. Эта совокупность собиралась членами Таллиннской творческой группы на протяжении всего времени ее

активной деятельности. Такого рода работа оказалась достаточно трудоемкой, однако она позволила существенно влиять на текущее и перспективное планирование (*если можно так выразиться относительно научной и любой иной творческой активности*) всей деятельности Группы. Действительно, работа потребовала сбора информации по проблематике группы во всех доступных как отечественных, так и зарубежных информационных источниках и, в первую очередь, в различного направления реферативных журналах. Собранная в процессе данной работы информация позволила Группе держать руку на пульсе развития данного направления современной кибернетики и своевременно реагировать на изменения его основных тенденций, методов исследования, а также приложений в различных областях естествознания. Наконец, вышеуказанная учетно-статистическая работа позволила собрать уникальную литературу и материалы для многих обзорных работ по данному важному разделу современной математической кибернетики.

**Таблица 1. Распределение научных публикаций ТТГ по типу и месту издания за 1970-1999 годы ее творческой активности**

| *Научные публикации по клеточным автоматам и их приложениям* | | *Количество публикаций* | *Средние затраты в д/нч.л.* | *% к групповому итогу* | *% к общему итогу* |
|---|---|---|---|---|---|
| *По типу* | *По месту* | | | | |
| *Монографии* | *Эстония* | *2* | *11.7* | *28.6* | *1.1* |
| | *Украина* | *1* | *10.3* | *14.2* | *0.6* |
| | *Россия* | *2* | *11.4* | *28.6* | *1.1* |
| | *Белоруссия* | *2* | *11.6* | *28.6* | *1.1* |
| | *Всего* | *7* | *11.25* | *100* | *3.9* |
| *Книги* | *Эстония* | *7* | *12.5* | *21.2* | *3.9* |
| | *Украина* | *9* | *15.5* | *27.3* | *5.0* |
| | *Россия* | *8* | *14.6* | *24.2* | *4.4* |
| | *Белоруссия* | *8* | *16.2* | *24.2* | *4.4* |
| | *Литва* | *1* | *9.2* | *3.1* | *0.6* |
| | *Всего* | *33* | *13.6* | *100* | *18.3* |
| *Сборники* | *Эстония* | *4* | *17.3* | *80* | *2.2* |
| | *Украина* | *1* | *17.7* | *20* | *0.6* |
| | *Всего* | *5* | *17.5* | *100* | *2.8* |
| *Научные отчеты* | *Эстония* | *6* | *37.6* | *75* | *3.2* |
| | *Россия* | *1* | *37.8* | *12.5* | *0.6* |
| | *За рубежом* | *1* | *40.2* | *12.5* | *0.6* |
| | *Всего* | *8* | *38.5* | *100* | *4.4* |
| *Научные статьи* | *Эстония* | *45* | *35.3* | *35.3* | *25.0* |
| | *СССР* | *12* | *35.7* | *9.5* | *6.7* |
| | *Белоруссия* | *12* | *34.8* | *9.5* | *6.7* |
| | *Литва* | *10* | *34.6* | *7.9* | *5.6* |
| | *За рубежом* | *48* | *36.4* | *37.8* | *26.7* |
| | *Всего* | *127* | *35.4* | *100* | *70.6* |
| *Общее количество:* | | *180* | *24.4* | *100* | *100* |

*A*

*B*

Из табл. **1,** в частности, следует, что из общего числа публикаций **ТТГ** почти треть (29.4%) приходится на большие издания (*монографии*, *книги*, *сборники статей и научные отчеты*), а среди них чуть больше трех четвертей (75.5%) – на книги. Среди научных статей больше трети (37.8%) приходится на зарубежные издания. В данном примере проведена *группировка* всех публикаций **ТТГ** за период 1970-1999 годы по двум признакам: тип (*монография*, *книга*, *сборник*, *отчет*, *статья*) и место публикации (*Эстония*, *Украина*, *Россия*, *Белоруссия*, *Литва*, *СССР* и *за рубежом*). Число групп в *типологической группировке* зависит от числа действительно имеющихся типов, при этом группировочные признаки могут быть как *качественными*, так и *количественными*. Как правило, значение *типологических группировок* определяется *сложностью* изучаемых объектов.

***Структурная*** группировка выявляет состав (*структуру*) изучаемой совокупности. Она имеет большое значение при изучении *предприятий* по отраслям производства, величине основных фондов, уровню механизации, числу рабочих и др. *Структурные группировки* находят весьма широкое применение и при анализе выполнения производственных планов. Структурная группировка может вестись и по *качественному*, и по *количественному* признаку. Примером такой структурной группировки служит разбиение научных статей по отечественным и зарубежным изданиям (табл. **1**). Следует отметить, что само подразделение группировок на *типологические* и *структурные* во многих случаях достаточно условно, что и иллюстрирует табл. **1,** ибо тип и структура тесно связаны между собой.

***Аналитическая*** группировка определяет взаимосвязи между двумя или более признаками совокупности. В статистике *зависимые* признаки называют ***результативными,*** а оказывающие на них влияние – ***факторными***. Признаки обоих типов могут быть как количественными, так и качественными (*атрибутивными*). Примером *аналитической группировки* может служить табл. **1,** в которой *факторный признак* (*тип публикации*) определяет изменение *результативного* признака (*средние затраты на подготовку рукописи к изданию в дн/пл=день/печ.лист*; *при расчете показателя учитывалось время от начала работы по теме до сдачи ее в издательство*). Видно, что показатель средних затрат растет от группы к группе. Качественный анализ этой (*в какой-то мере*, *на первый взгляд*, *не совсем ожидаемой*) зависимости показывает, что средние затраты по первым трем группам изданий (*монографии*, *книги*, *сборники*) сопоставимы и по *первой* группе *минимальны*, достигая *максимума* для третьей группы. Объясняется эта ситуация следующими основными обстоятельствами:

- ***монографии*** *писались, в основном, на основе готовых научных результатов, в большинстве своем ранее опубликованных, и носили, в основном, обобщающий и систематизирующий характер*
- ***книги,*** *в основном, посвящены программным средствам для различного типа и класса ЭВМ; их создание в целом ряде случаев шло параллельно с освоением, апробацией и адаптацией этих средств, что требовало серьезных проработок*
- ***сборники*** *же наряду с готовыми материалами включали и новые разработки научно-прикладного характера, требующие значительных усилий*
- ***статьи*** *в отечественных и зарубежных изданиях содержали, в основном, оригинальные теоретические результаты, требующие глубокой проработки с привлечением ряда разделов современной математики, кибернетики, теоретической биологии, вычислительных и других наук.*

Эти обстоятельства и обусловили более чем двухкратный рост средних затрат по **4-**й и **5-**й группам относительно *аналогичного среднего* показателя по *первым* трем группам. Естественно, результаты данного анализа отражают только *творческую* активность **ТТГ** и не могут служить основой для обобщающих выводов научной активности в целом, но методика выявления тенденции заслуживает определенного внимания.

По числу группировочных признаков все группировки делятся на *простые*, *комбинированные* и *многомерные*. Группировка совокупности по одному, более одного и большому количеству признаков называется соответственно ***простой, комбинированной*** и ***многомерной***. В табл. **1** приведен пример комбинированной группировки публикаций **ТТГ** по двум признакам – *типу* и *месту* публикаций, что позволяет анализировать *влияние* одного и другого признаков. Использование нескольких группировочных признаков требует значительного объема совокупности, так как в противном случае отдельные группы могут содержать недостаточное число единиц для получения обоснованных выводов. Поэтому современная статистическая практика оперирует не более, чем с **3-4** группами.

В настоящее время весьма большое внимание уделяется вопросам *многомерной* группировки. Задача группировки по любому числу признаков может быть решена одним из методов статистической теории распознавания образов – ***кластерным анализом*** [110]. Разработка данного метода восходит к 60-м годам и тесно связана с использованием ЭВМ и, в первую очередь, с классом **ПК**. Суть данного метода состоит в следующем. Группировка единиц производится не последовательно по отдельным признакам, а одновременно по их набору, образующему *признаковое пространство*. Если для совокупности имеется **N** признаков, то каждая единица рассматривается как точка в **N**-мерном пространстве и задача сводится к выделению сгущений точек (*кластеров*) в этом пространстве на основе их геометрической близости или другой меры сходства (*близости*). *Многомерные* группировки позволяют решать целый ряд весьма важных задач экономико-статистического исследования, однако из-за ряда нерешенных вопросов метода распознавания образов следует критически относится к получаемым на его основе результатам и их последующей трактовке. Основные элементы метода группировок основаны на мере ассоциации и стратегии, очень широко принятой при формировании групп переменных из ассоциативной таблицы (*см. например*, [193]).

Альтернативный подход состоит в том, чтобы рассматривать набор ассоциаций между парами переменных в качестве обеспечения мер расстояний между переменными и пытаться изображать набор расстояний геометрически в 2- или 3-мерном пространмстве, не вводя слишком большие искажения в относительные расстояния. Методы, которые используются в рамках данного подхода, называют методами *классификации*, *координатным анализом* [195], и *многомерным шкалированием* [194]. Когда количество переменных является маленьким, есть возможности того, чтобы определять и исследовать разные статистические модели, которые изображают характер взаимосвязей между полным набором переменных. Систематическое развитие в статистическом подходе стало возможным в связи с высокой скоростью обработки вычислительными средствами, обсуждению которых посвящена последняя глава книги.

По исходным данным группировки делятся на ***первичные*** и ***вторичные***. В *первом* случае группировка производится на основе первичной статистической информации, во *втором* – на основе уже имеющейся группировки. Если пример в табл. **1** считать *результатом первичной* группировки, то в качестве *вторичной* можно определить разбиение всех публикаций на две группы: в *непериодической* (**А**) и *периодической* (**В**) печати. Поэтому первые 3 группы *первичной* группировки составят **А**-группу вторичной, а последние 2 первичной – **В**-группу вторичной. В результате *вторичной* группировки более четко выступают различия в средних затратах на подготовку изданий обоих типов. Следует иметь в виду, что в результате перегруппировок необходимо осторожно оперировать со *средними средних*. Об этом будет идти речь несколько ниже. *Вторичная* группировка производится для увеличения или уменьшения числа ранее образованных групп, в основном, двумя способами: изменением интервала первичной группировки или по удельному весу групп в общем итоге. Метод вторичной группировки позволяет проводить многоаспектный анализ статистических данных, полученных в результате статистического наблюдения.

## 4.3. Интервальные группировки и классификации

При группировках по количественным признакам возникают вопросы о количестве групп и величине интервала. При их решении следует учитывать два момента: (**1**) *диапазон изменения значений признака* и (**2**) *объем изучаемой совокупности*. Как правило, в большей совокупности образуется большее количество групп. Однако, если распределение единиц совокупности по *группировочному* признаку близко к *нормальному* и в группах применяются равные интервалы, то ориентиром при определении числа групп (**n**) в зависимости от объема (**N**) совокупности может служить следующая формула Стерджесса:

$$\mathbf{N = 1 + 1.4427*ln(N)} \qquad (30)$$

Так, при **N**=10 и **N**=100 получаем соответственно **n1** = [1 + 3.32] = **4** и **n2** = [1 + 6.64] = **8** групп. Под интервалом (**Int**) понимается разница между *максимальным* и *минимальным* значениями признака **Xj** переменной **X** в кажждой группе, а именно: **Intj = (*max*{Xj} - *min*{Xj}); j=1 .. n**. При равных интервалах все значения **Intj** равны, а именно: **Intj = (*max*{Xj} - *min*{Xj})/n,** где **Xj** есть значения признака в совокупности и **n** – количество групп; в противном же случае используются *неравные* интервалы. *Равные* интервалы используются, когда изменение количественного признака внутри совокупности происходит равномерно. *Неравные* интервалы, как правило, используются тогда, когда варьирование признака осуществляется неравномерно и/или в широких диапазонах. В свою очередь, *неравные интервалы* могут быть как *возрастающими*, так и *убывающими*.

В группировках, имеющих целью отобразить качественное своеобразие групп, применяются *специальные* интервалы. В этом случае каждая группа имеет особое содержание, а граница интервала определяет переход от одного качества к другому. Специальные интервалы и их размеры определяются особенностями объекта исследования. Приведем примеры на все три типа интервальных группировок. *Равные* – количество публикаций за пятилетки: 1970-1974, 1975-1979,1980-1984,1985-1989,1990-1994, 995-1999; *неравные* – количество рабочих по проценту выполнения норм выработки, а именно: **>** 200, 200-150, 149-120, 119-110, 109-100, 99-95, 94-90, **<** 90; *специальные* – возрастное распределение мужского населения по отношению к трудовой активности: **<** 15, 16-18, 19-59, 60-69, **>** 69.

Задача всестороннего анализа социально-экономических явлений требует создания системы группировок, число которых в настоящее время достаточно велико. Среди них вполне можно отметить группировки: населения, городов, затрат на производство, в статистике труда, предприятий промышленности, полезных ископаемых и другие [96, 115]. Создание системы группировок по многим признакам должно отвечать целому ряду общих методологических требований, подчиняющихся некоторым логическим и формальным критериям [92].

***Классификация*** – единообразная группировка явлений и объектов по классам и группам, имеющая общеметодологическое значение и предусматривающая введение общих разделов и частных подразделов. От обычной *группировки классификация* отличается более детальным и развернутым разбиением совокупности объектов, перечень которых рассматривается как статистический стандарт, утверждаемый, как правило, в качестве национального либо международного стандарта. В каждой стране действуют национальные классификации, которые разрабатываются и утверждаются центральными статистическими органами. Например, в СССР были созданы классификации: отраслей народного хозяйства, профессий и занятий, грузов, библиотек, основных фондов, научных исследований и разработок, и др. [96]. Другим примером является **УДК** – универсальная десятичная классификация разделов современной науки {**UDK** (*Brussels classification*, *Universal Decimal classification*)}; ту же роль для

*физико-математических наук* выполняет международная *классификация* **AMS** Американского математического общества. Следует иметь в виду, что в отличие от *классификации группировка* может быть произведена только для целей конкретного исследования, для выяснения в нем одного или нескольких отдельных вопросов. Тогда как *классификация* – регламентирующий документ длительного пользования с весьма широким диапазоном применения. Например, классификация научных исследований и разработок четко дифференцирует всю научную активность на три класса: ***фундаментальные исследования, прикладные исследования*** и ***разработки***. Именно в соответствии с этими классами производится градация всей научно-исследовательской деятельности как творческих коллективов, так и отдельных *исследователей* в различных областях современной науки.

Наряду с классификациями в статистике весьма широко используются *номенклатуры*, под которыми понимается стандартный перечень объектов и групп, входящих в определенную классификацию. Следует иметь в виду, что классификация представляет собой более глубокое, научно обоснованное деление групп, а *номенклатура* – более узкое, техническое перечисление различных объектов. Наличие разных классификаций как национальных, так и международных позволяет сводить и сопоставлять различного рода данные, дает хорошую возможность обобщать статистические данные. Подробное рассмотрение вопросов *различных* классификаций и номенклатур входит в курсы экономической и отраслевых статистик народного хозяйства страны. Здесь лишь отметим, что задача создания различного рода классификаторов является достаточно сложной и трудоемкой. Современные проблемы анализа данных и их классификации достаточно подробно рассматриваются в книге [297].

## 4.4. Табличное представление статистических данных

*Табличный* способ представления информации чрезвычайно распространен во всех сферах человеческой деятельности и имеет длинную историю (*таблицы логарифмов, случайных чисел, деления, умножения, вычисления корней, интегралов, процентов и др.*). Однако далеко не каждая таблица является статистической. ***Статистическая*** таблица в отличие от нестатистических отличается тем, что в ней дается сводная характеристика статистической совокупности, подводится *один* или *несколько* итогов. *Статистические таблицы* (*ниже просто* ***статтаблицы***) – наиболее эффективная форма представления результатов сводки. В таблице наиболее наглядно проявляется связь между признаками изучаемого явления, процесса или объекта. Первые статтаблицы были использованы русским географом и статистиком И. Кирилловым в 1727 для описания отдельных областей и всего Российского государства в целом. В западной статистике изобретение табличного способа представления информации приписывается датчанину И. Андерсену. Однако его работа вышла в свет только в 1741 г., а используемые в ней таблицы были существенно проще таблиц И.К. Кириллова.

*Составление макетов* статтаблиц – *важнейшее* условие *планирования* разработки статистических материалов. Таблица состоит из трех основных компонент: *общий заголовок таблицы* (*General header of the table*), *заголовок подлежащего* (*Subject header*) и *заголовок сказуемого* (*Predicate header*). На рис. **7** представлена общая схема статистической таблицы.

***Общий заголовок*** (*General header*) кратко информирует о содержании статтаблицы, к какому месту и времени относятся ее данные. Кроме общего в типичной таблице имеются ***заголовки подлежащего*** (*Subject header*; боковые, строчные) и ***сказуемого*** (*Predicate header*; верхние, графовые). *Подлежащее* статтаблицы определяет перечень отдельных единиц или групп характеризуемого явления или объекта. Как правило, подлежащее располагается в таблице слева и определяет содержание ее строк (*rows*). *Сказуемое* определяет перечень признаков, характеризующих подлежащее, записывается вверху таблицы и определяет содержание ее

граф (*столбцов*, *columns*). *Клетки* (*Cell*), образуемые на пересечении строк-подлежащим и графами-сказуемым, содержат числовые данные, характеризующие те или иные аспекты исследуемого явления. Размер таблицы определяется как произведение (**mxn**), где **m** и **n** – соответственно число строк и столбцов таблицы. Для случая табл. **1** получаем значение (**mxn**) = **24x4**. Каждая таблица при необходимости может снабжаться *специальными* и/или *общими пояснениями* (*general and/or special notes*) о содержании приводимых в ней данных, методике их получения, единицах их измерения и др. В таблице отдельно выделяются *итоговые строка* (*balance row*) и *столбец* (*balance column*).

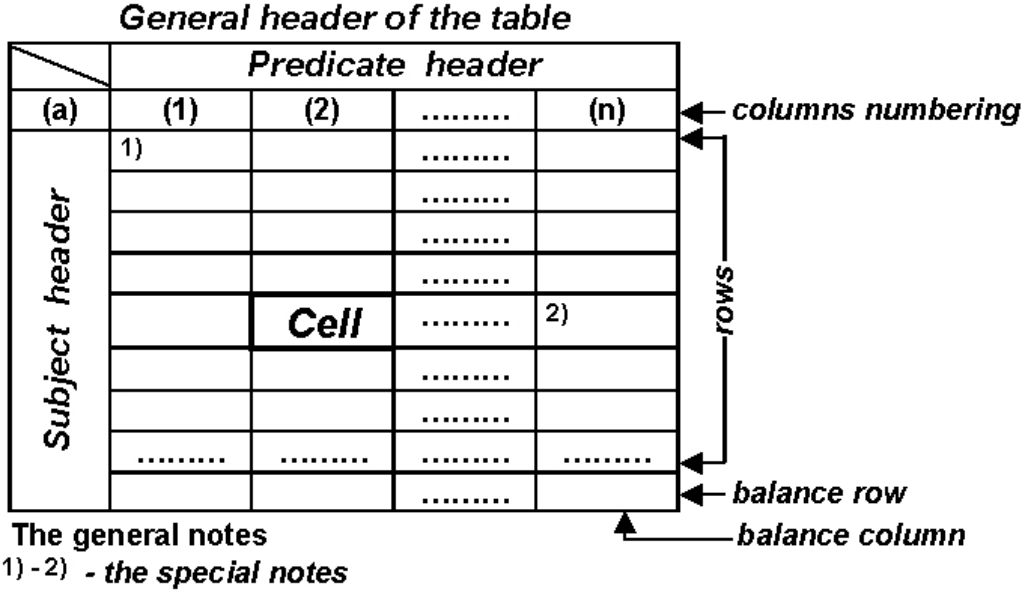



**Рис. 7. Общая схема-макет статистической таблицы**

В зависимости от разработки подлежащего или группировки единиц в нем выделяются три типа таблиц, а именно: *простые*, *групповые* и *комбинационные*. ***Простая*** (*или перечневая*) таблица содержит сводные показатели или перечень отдельных элементов без разбиения совокупности на группы. Так, нижеследующая табл. **2** имеет *простой* тип, характеризуя постоянное население Эстонии по полу и возрасту.

**Таблица 2. Постоянные жители Эстонии по полу и возрасту (на 1.01.1999)**

| *Возрастные группы* | *Население* | | | *Разность* |
|---|---|---|---|---|
| | *Всего* | *мужчины* | *женщины* | **(4) = (2) - (3)** |
| **(a)** | **(1)** | **(2)** | **(3)** | **(4)** |
| **0 - 4** | **64612** | **33174** | **31438** | **1736** |
| **5 - 9** | **92675** | **47648** | **45027** | **2621** |
| **10 - 14** | **111547** | **56662** | **54885** | **1677** |
| **15 - 19** | **106233** | **53889** | **52344** | **1545** |
| **20 - 24** | **102895** | **52051** | **50844** | **1207** |
| **25 - 29** | **106765** | **55557** | **51208** | **4349** |
| **30 - 34** | **97420** | **49218** | **48202** | **1016** |

| | | | | |
|---|---|---|---|---|
| **35 - 39** | **106153** | **52229** | **53924** | **-1695** |
| **40 - 44** | **105595** | **50705** | **54890** | **-4185** |
| **45 - 49** | **99252** | **46487** | **52765** | **-6278** |
| **50 - 54** | **81711** | **37165** | **44546** | **-7381** |
| **55 - 59** | **82621** | **36471** | **46150** | **-9679** |
| **60 - 64** | **81286** | **34001** | **47285** | **-13284** |
| **65 - 69** | **72537** | **28243** | **44294** | **-16051** |
| **70 - 74** | **60374** | **20139** | **40235** | **-20096** |
| **75 - 79** | **36412** | **10065** | **26347** | **-16282** |
| **80 - 84** | **19815** | **5089** | **14726** | **-9637** |
| **85 -** | **17677** | **3883** | **13794** | **-9911** |
| ***Итого:*** | **1445580** | **672676** | **772904** | **-100328** |

*Простая* табл. **2** содержит в качестве *подлежащего* список возрастных групп, а в качестве *сказуемого* – четыре переменные (*Общее количество*, *мужчины*, *женщины* и *Разность*), которые характеризуют предмет. Простые таблицы в статистических публикациях имеют самое *широкое* применение и используются намного чаще, чем таблицы *групповые* и *комбинационные* [29, 32, 74, 146, 188, 199]. Основное назначение такого типа таблиц состоит в четком описании исследуемого объекта с целью информации читателя статистической публикации.

***Групповая*** таблица в подлежащем содержит группировку единиц исследуемой совокупности только по одному признаку. Примером такого типа таблиц может служить табл. **1,** если в ее подлежащем оставить только группировку публикаций **ТТГ** по их типу. Следовательно, в отличие от *простой* таблицы каждая строка *групповой* таблицы содержит уже некоторую группу совокупности, а не ее отдельную единицу. *Групповые* таблицы имеют весьма важное познавательное значение и используются для выявления и характеристики различных типов объекта, его структуры, взаимосвязей между его отдельными компонентами.

***Комбинационная*** таблица в подлежащем содержит группировку единиц совокупности по двум или более признакам. В такой таблице группы, образованные по первому признаку, расчленяются на подгруппы по второму, третьему и т.д. признакам. Это дает возможность рассматривать влияние на признаки сказуемого не одного, а более факторов подлежащего. Типичным примером таблиц такого типа является табл. **1,** в которой подлежащее содержит группировку по двум признакам (*типу* и *месту публикации*). В качестве довольно полезного упражнения читателю рекомендуется провести по ней анализ публикаций **ТТГ** по каждому из признаков подлежащего и по их совокупности. *Комбинационная* таблица по **К** признакам не может быть заменена эквивалентным ей набором **К** *групповых* таблиц, ибо отсутствует возможность анализа изменения признаков сказуемого в зависимости от совокупности признаков подлежащего. Однако в целом ряде случаев такое разбиение может быть весьма полезным.

Статистическая практика показывает, что группировки по трем и более признакам применяются довольно редко, ибо число образуемых подгрупп растет в геометрической прогрессии и таблицы становятся плохо обозримыми. В *комбинационных* таблицах нельзя произвольно менять место данного признака в их комбинации. Их следует расставлять либо по важности, либо по порядку изучения. Сводку результатов наблюдения лучше всего

начинать с построения *комбинационных таблиц*, которые при *необходимости* можно разбивать на совокупности *групповых*. Из *групповых* же *комбинационные* таблицы составлять нельзя.

По способу разработки сказуемого таблицы делятся на *простые* и *комбинированные*. В случае простой разработки сказуемое содержит столько признаков, сколько их регистрировалось по единице совокупности в процессе наблюдения. При комбинированной каждая группа *подлежащего* может характеризоваться *любым числом признаков сказуемого*. При *комбинированной* разработке сказуемого появляется возможность проведения более глубокого статистического анализа, с другой стороны, большое число граф делает таблицу плохо обозримой.

Наряду с рассмотренными выше основными тремя типами статтаблиц следует упомянуть о ***рабочих*** и ***разработочных*** таблицах. Первые составляются в процессе первичной сводки результатов наблюдения, имеют, как правило, более широкие масштабы графления и допускают производить в них различного рода вычисления производных величин (*средних*, *отклонений*, *относительных величин и др.*). *Разработочная* таблица составляется при первичной сводке результатов наблюдения и содержит, как правило, более детальные группировки, чем окончательная таблица. Детальная группировка позволяет легко на ее основе создавать требуемые группировки. Иногда данные разработочной таблицы могут служить основой углубленного исследования отображаемого ею объекта. Среди специальных статтаблиц можно отметить таблицы: *рождаемости*, *смертности*, *трудовой занятости* и другие [115, 199]. В научных исследованиях табличное представление информации и на его основе анализ тех или иных явлений и процессов также имеет весьма большое значение. При таком подходе создаются, порой, весьма оригинальные *формы* и *типы* таблиц, отражающих суть *исследуемого* явления и позволяющие значительно нагляднее отображать его специфику.

Для обеспечения потребностей в табличном представлении статистических данных, современное программное обеспечение (*математические и статистические пакеты для* **ПК,** *прежде всего*) располагает рядом эффективных средств для построения табличных объектов различных видов и их обработки. В частности такое средство как *Microsoft Excel*, будучи частью пакета *Microsoft Office*, позволяет легко создавать, редактировать, обрабатывать и выводить табличные объекты на основе пользовательских данных. Начальные данные для таблицы могут быть введены непосредственно с консоли компьютера или из файлов данных, предварительно подготовленных в среде популярных средств таких как: *Access*, *dBase*, *FoxPro*, *Paradox*, *Lotus 1-2-3*, или из файлов данных *ASCII*-типа. Все это существенно облегчает задачу табличного представления статистических данных любого характера и назначения. Кроме того, для ряда программных средств именно табличное представление статистических данных допустимо для реализации их статистического анализа (*например*, *для Microsoft Excel*). Данный вопрос будет рассматриваться более детально несколько ниже.

## 4.5. Статистические ряды распределения

Построение ***рядов распределения*** является составной частью сводной обработки данных наблюдения. ***Статистическими рядами*** называются ряды распределения, полученные в результате наблюдения. Ряды распределения характеризуют распределение единиц совокупности на группы по какому-нибудь варьирующему признаку: *атрибутивному* или *количественному*. В первом случае ряды распределения называются ***атрибутивными,*** во втором – ***вариационными*** (*упорядоченные выборки*). Отдельную группу атрибутивных статистических рядов составляют ***географические,*** построенные по географическому признаку (*например*, *распределение населения по континентам*, *климатическим зонам*, *расам и др.*). Другим примером может служить распределение населения государства по районам и областям, полу, образованию, профессиональной ориентации и др.

***Вариационный ряд*** (**ВР**) дает *распределение* единиц совокупности по значениям варьирующего признака и определяется двумя элементами: *вариантами* и их *частотами*. ***Вариантами*** называются отдельные значения группировочного признака, принимаемые им в **ВР**. ***Частотой*** данного варианта называется количество его в совокупности, тогда как *частость* варианта есть его частота в долях или процентах к итогу и часто используется наравне с частотой. Сумма всех частостей в пределах совокупности равна **1** (*доля*) или **100%** (*процент*). Класс всех **ВР** делится на *дискретные* и *интервальные*. В *дискретном* **ВР** вариант принимает дискретные значения, как правило, целочисленные (*количество человек в семье, тарифный разряд, число публикаций и др.*). В случае *интервального* **ВР** значения варианта даются в виде *интервалов*, которые получаются в результате *группировки* данных наблюдения. *Интервальные* **ВР** строятся по *непрерывному* (*возраст человека, зарплата, объем публикации, прибыль и др.*) или *дискретному* (*число студентов, публикаций и др.*) принципу, если дискретная вариация имеет широкие пределы.

*Интервальные* **ВР** имеют интервалы либо строго *определенные*, либо *нечеткие*. Примером интервального **ВР** с четко определенными интервалами является ряд, определяющий распределение семей по количеству их членов: 1-2, 3-4, 5-6, 7-8, 9 и более. В случае таких **ВР** каждое значение варианта относится строго к одному интервалу. В случае *нечеткого* интервала необходимо уточнять, к какому интервалу относятся значения признака. Обычно они относятся к тому интервалу, для которого они совпадают с его нижней границей. В случае наличия для **ВР** *незакрытых* интервалов типа "*до* **n**" и/или "*больше* **m**" совокупность содержит небольшое число единиц, отличающихся по значению от основного ее массива. Для определения различных статистических показателей (*средняя, отклонение, дисперсия и др.*) *интервальный* **ВР** следует преобразовывать (*условно*) в дискретный, для чего необходимо:

- **ВР** *с незакрытыми интервалами преобразуются в* **ВР** *с закрытыми интервалами*; *при равных интервалах, если ничего не известно о характере граничных интервалов, они полагаются равными по длине остальным*
- *интервальный* **ВР** *преобразуется в дискретный путем замены интервалов средними арифметическими их пределов*.

Проще всего ряды распределения анализировать при помощи их графического изображения, позволяющего судить о форме распределения данных. Наглядное представление о характере изменения частот вариационного ряда дают такие графики как: (**1**) ***полигон*** распределения частот, (**2**) ***гистограмма***, (**3**) ***кумулята*** и (**4**) ***огива***.

Для наглядности *дискретные* и *интервальные* **ВР** графически изображаются соответственно ***многоугольником*** (*полигоном*) и ***гистограммой*** распределения. Для **ВР** можно строить еще два вида графиков: ***кумулянты*** (*кривые сумм*) накопленных частот и ***кривые концентрации*** (*кривые Лоренца*). ***Полигон*** представляет собой ломаную линию в системе координат. Ее конфигурация отражает специфику распределения значений варьирующего признака. Наиболее показательно совмещение двух или нескольких полигонов распределения одного и того же признака в одной системе координат. ***Гистограмма*** представляет собой график в декартовой системе координат, отражающий по оси ординат **Y** *частоту* накопления величин в классе, а по оси абсцисс **X** – *границы* классов. *Кумулятивная кривая* (***кумулята***) – ломаная, составленная по *последовательно* суммированным, т.е. накопленным частотам или относительным частотам. Если при изображении *кумуляты* поменять оси координат местами, то получится ***огива***. При построении кумулятивной кривой дискретного признака на ось абсцисс наносятся значения признака, а ординатами служат нарастающие итоги частот. *Кривой концентрации* или ***кривой Лоренца*** называют кривую относительной концентрации суммарного значения признака. *Графическое* представление *вариационного ряда* может быть дополнено *графическими объектами* – линиями, стрелками, полями с надписями,

различными геометрическими фигурами. При этом, можно определять шрифт, узор, цвет и тип линии, месторасположение объектов по отношению к другим объектам и рабочему полю графика. Подобные графические объекты весьма легко создаются с помощью команд, например, в электронных таблицах *Microsoft Excel*. Графическое представление результатов измерений не только существенно облегчает анализ и выявление скрытых закономерностей, но и позволяет правильно выбрать последующие статистические характеристики и методы анализа.

С целью пояснения и иллюстрации указанных типов графиков приведем первоначальную обработку интервального **ВР** с незакрытыми граничными интервалами. Имеются данные (табл. **3**) по объему и месту периодических публикаций **ТТГ** [23, 29, 127, 191]. Естественно, публикации, составившие **ВР,** различны по своему характеру, поэтому их совокупность нельзя рассматривать *однородной*. Но в отношении рассматриваемого признака (*объема публикации*) данная совокупность становится уже достаточно однородной.

**Таблица 3. Распределение периодических публикаций ТТГ по месту их издания и объему (1970 – 1999)**

| *Объем статей в страницах* | *Замкнутые интервалы* | *Зарубежные издания* **(А)** | | | | *Отечественные издания* **(В)** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *Количество* | *Общий объем (в страницах)* | *Доли объемов (%)* | *Доли объемов (%)* | *Количество* | *Общий объем (в страницах)* | *Доли количеств (%)* | *Доли объемов (%)* |
| | | | | *С нарастанием объема* | | | | *С нарастанием объема* | |
| **< 10** | **1 - 10** | **14** | **81** | **29.1** | **10.1** | **38** | **204** | **48.1** | **11.7** |
| **10 - 20** | **10 - 20** | **20** | **294** | **70.8** | **47.1** | **18** | **227** | **70.9** | **24.6** |
| **20 - 30** | **20 - 30** | **10** | **249** | **91.6** | **78.3** | **6** | **140** | **78.5** | **32.6** |
| **30 - 40** | **30 - 40** | **2** | **68** | **95.8** | **86.8** | **5** | **167** | **84.8** | **42.1** |
| **40 - 50** | **40 - 50** | **1** | **47** | **97.9** | **92.7** | **4** | **170** | **89.9** | **51.8** |
| **50 - 60** | **50 - 60** | **1** | **58** | **100** | **100** | **1** | **58** | **91.1** | **55.1** |
| **> 60** | **> 60** | **0** | **0** | **100** | **100** | **7** | **788** | **100** | **100** |
| | ***Итого*** | **48** | **797** | **-** | **-** | **79** | **1754** | **-** | **-** |

Прежде всего, по формуле (30) определяем число равных интервалов следующим образом:

$$n1 = [1+1.4427*\ln(48)] = 7, \quad n2 = [1+1.4427*\ln(79)] = 7$$

т.е. **n = n1 = n2 = 7,** и преобразуем интервалы **А** и **В** полученных **ВР** в закрытые (*табл.* **3,** *графа* **2**). Затем строим для обоих рядов (*рис.* **8, а**) *гистограммы* (*сплошными линиями для* **ВР В** и *пунктирными для* **ВР А**). Гистограмма строится в *декартовой* системе координат; по оси *абсцисс* откладываются отрезки, изображающие интервалы значений варьирующего признака. На них, как на основаниях, строятся прямоугольники, высота которых при *равных* интервалах соответствует *частотам* или *частостям*, а при *неравных* – *плотностям* распределения соответствующих интервалов. Получаем ступенчатую фигуру в виде примыкающих друг к другу прямоугольников, площади которых пропорциональны частотам (*частостям*).

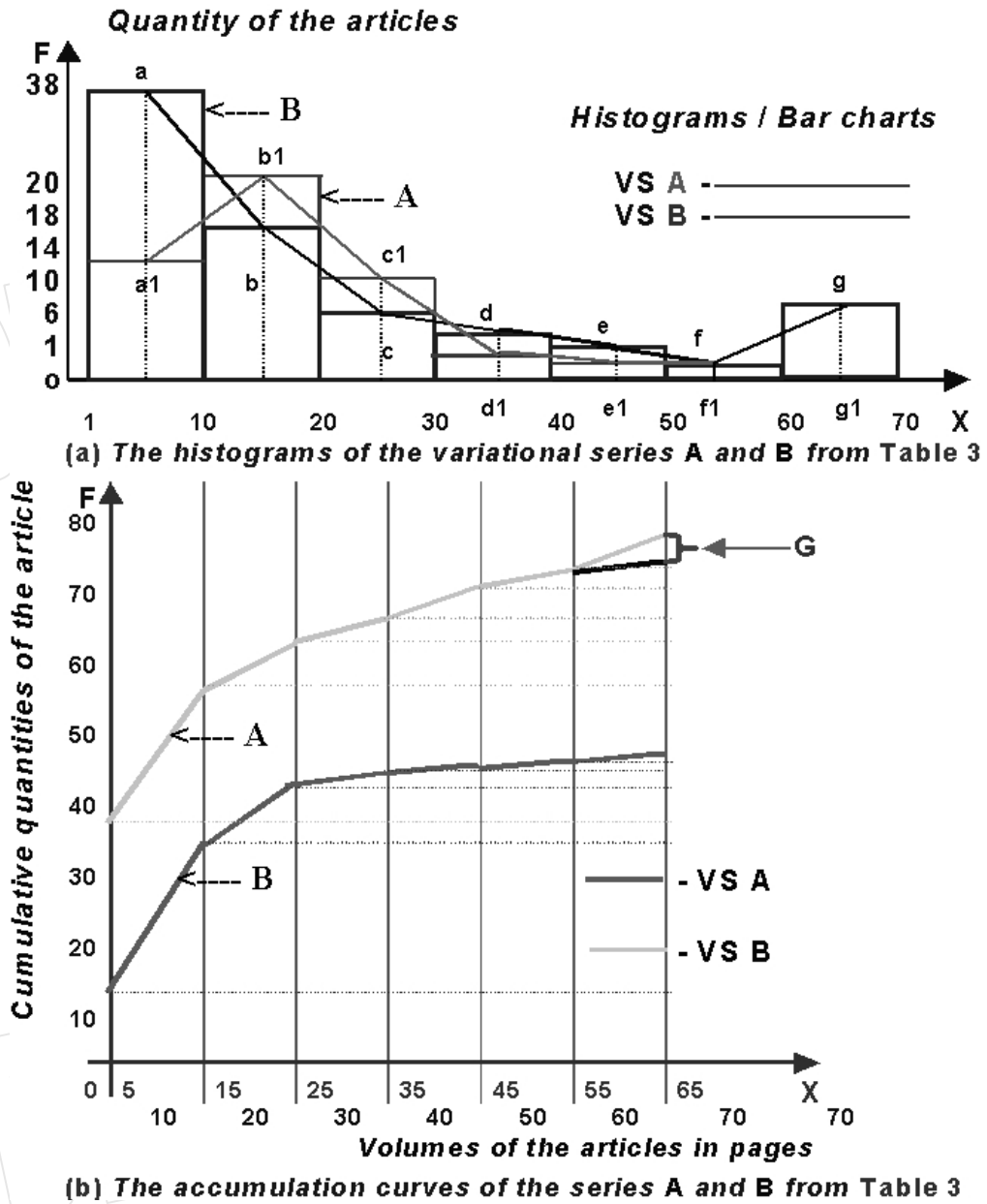



**Рис. 8. Гистограммы, полигоны и интегральные кривые распределения**

Из гистограммы легко получить ***полигон распределения***, для чего необходимо соединить ломаной линией середины верхних сторон прямоугольников (рис. **8, а;** для **ВР В** – сплошная ломаная линия **abcdefg,** для **А** – ломаная **a1b1c1d1e1f1g1**). Для дискретных **ВР** *полигоны* распределений строятся непосредственно. При увеличении числа наблюдений из одной и той же совокупности увеличивается число групп интервального **ВР,** что приводит к уменьшению величины интервала. При этом, число сторон соответствующего полигона распределения будет расти и его ломаная линия будет стремиться к плавной *кривой распределения*. Следует отметить, что *кривая распределения* отражает теоретическое распределение, получаемое при полном погашении всех случайных причин, затеняющих основную закономерность исследуемого явления как такового.

В ряде случаев для изображения **ВР** используется *кумулятивная кривая* (***кумулята***), для чего подсчитываются по интервалам накопленные частоты или частости. Накопленные частоты

показывают, сколько единиц совокупности имеют значения признака не большие, чем рассматриваемое значение. На рис. **8** (**b**) представлены кумуляты (*ломаные линии*) соответственно для **BP A** и **BP B.** Для получения кумулят накопленные частоты (*частости*) наносятся на график в виде перпендикуляров (*длиной*, *равной частоте или частости*) к **X**-оси в точках, отмечающих полусуммы интервалов. Затем концы перпендикуляров соединяются прямыми. Кумуляты удобны при сравнении различных статистических рядов, а также в экономических исследованиях, в частности, при анализе концентрации производства. В частности, из нашего конкретного примера следует, что процесс концентрации объема публикаций *зарубежных* и *отечественных* (рис. **8, b;** *кумуляты* **A** *и* **B**) происходил, практически, одинаково, кроме интервала (60-70), где "*всплеск*" **G** объясняется публикацией достаточно объемных (*более* 100 *стр*.) наших статей в сборниках трудов [24-27].

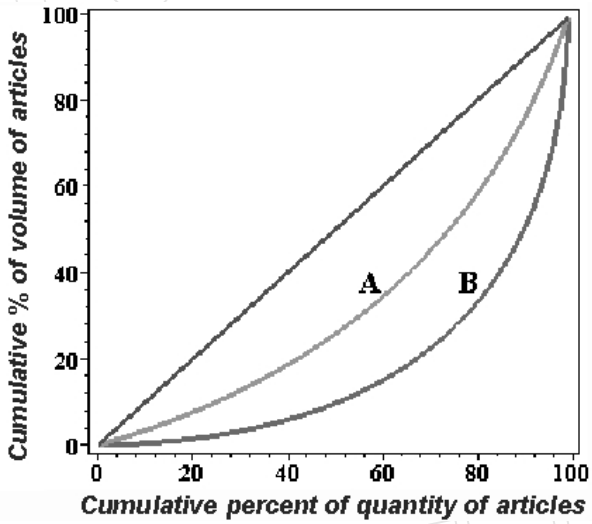



**Рис. 8 (c). Кривые Лоренца для объемов и количества статей ТТГ**

*Кривая Лоренца* графически характеризует уровень концентрации отдельных единиц совокупности по группам. Для ее построения на оси координат наносится процентная масштабная шкала от **0** до **100%** (рис. **8, c**). Затем для **BP A** вычисляем два показателя по каждому интервалу: проценты *количества* и *объема* статей к итоговому значению. После этого формируем их значения по интервалам ряда с нарастающим итогом (*табл*. **3,** *графы* **3** *и* **4**). Оси абсцисс и ординат графика отводим соответственно для кумулятивных (*cumulative*) процентов количества и объема статей – единиц **BP A.** Затем по значениям из граф **3** и **4** (*табл*. **3**) строим точки на графике и соединяем их плавной кривой **A.** Аналогично поступаем и для **BP B,** строя для него кривую Лоренца на том же графике, что позволяет проводить сравнительный визуальный анализ **BP A** и **B.** Прямая **Y = X** на графике (рис. **8, c**) определяет равномерное распределение – на один процент количества публикаций приходится один процент их объема. Следовательно, концентрация объема публикаций в группах по их количеству значительно выше для **BP B,** чем для **BP A,** т.е. концентрация зарубежных публикаций более близка к равномерной, что можно объяснить более целенаправленной ориентацией деятельности **ТТГ** на западного читателя.

Геометрически *степень близости* концентрации к *равномерной* определяет размер (**R**) площади фигуры, образованной кривой Лоренца и прямой **Y = X** – чем она меньше, тем степень *близости* больше. *Коэффициент Джини* **G=2*R** характеризует *меру концентрации*. Приведенная нами расчетная схема получения кривой Лоренца проста, однако логико-статистическое толкование модели, на которой она основывается, бывает весьма замысловато. По тематике

модели Лоренца опубликовано много работ по концентрации и подробнее с ней можно ознакомиться, например, в книгах [71 ,76, 78, 83]. Данная модель довольно интенсивно используется при моделировании целого ряда экономических процессов и явлений, а также в других теоретических разработках экономического характера.

## 4.6. Графическое представление статистических данных

***Графический** язык* является *особой* формой представления информации, научного мышления и обобщения. Графики, подобно другим искусственным языкам (*например*, *математическому*) имеют целый ряд положительных свойств, особенно в смысле лаконичности, однозначности, наглядности и интернациональности восприятия. Графические объекты дают возможность сопоставлять размеры совокупностей, создавать модели структур и динамики, размещения и связи явлений; они являются мощным средством анализа и прогнозирования.

Графические методы в статистике применяются уже около 200 лет и восходят к английскому экономисту У. Плейфейру, который выпустил в 1786 г. "*Коммерческий и политический атлас*"**.** Графическое изображение статистических данных прочно вошло в арсенал современных средств оформления результатов статистического анализа, их наглядного представления и обобщения. Особенно велико значение графиков в широкой пропаганде статистической информации *различного* назначения. Вопросы применения *графических методов* в статистике неоднократно обсуждались на многих международных статистических конференциях и конгрессах.

Графический язык статистики, не говоря уже о других сферах его применения, достаточно богат и разнообразен, что в рамках данной книги не позволяет рассмотреть его во всей его полноте. Поэтому мы ограничимся наиболее простыми и используемыми его компонентами. *Статистические графики*, в основном – геометрические двухмерные или трехмерные *условные* знаки, отражающие различные аспекты статистических совокупностей. Мы ограничимся двухмерным (*плоским*) случаем графиков, ибо с трехмерными (*объемными*) графиками лучше всего оперировать на видеомониторе ЭВМ. В графическом объекте можно выделить *пять* основных составных элементов (*при этом, часть из них может отсутствовать*):

1. ***поле графика*** *– пространство размещения различных знаков и символов, которое имеет определенные размеры и пропорции сторон*
2. ***смысловые знаки –*** *символы понятий, отражаемых на графике* (*точки, круги, линии и др.*)
3. ***пространственные ориентиры –*** *определяют размещение знаков в поле графика и зависят от принятой системы координат* (*декартова, полярная, сферическая, логарифмическая и др.*)
4. ***масштаб –*** *эталоны, отражающие размеры геометрических знаков; обычно размещаются вне поля графика*
5. ***комментарий –*** *словесное описание содержания графика и его компонент, если это требуется; как правило, сложные графики снабжаются достаточно подробной описательной частью для упрощения их последующего ипользования*.

Перед построением *статистического графика* определяются для него статистические данные и все его элементы. В предыдущем разделе рассмотрены четыре типа графиков *представления* и анализа статистических данных, а именно: (**1**) полигоны и (**2**) гистограммы распределения, (**3**) кумулятивные кривые и (**4**) кривые Лоренца. Здесь рассмотрим вопрос графического представления статистических данных несколько подробнее. Ввиду большого разнообразия графических объектов, отличающихся многими особенностями, возникает необходимость их классификации. В связи с *типом поля* графики делятся на *диаграммы* и *статистические карты*.

Диаграммы, в свою очередь, делятся на *диаграммы сравнения*, *структурные* и *динамические*, а также *графики связи*.

***Статистические карты*** можно разделить на *картограммы*, *картодиаграммы* и *центрограммы*. *Картограмма* иллюстрирует содержание статистических таблиц, подлежащим которых является административно-географическое деление совокупности. Существует много разновидностей картограмм (*фоновая*, *точечная*, *грузопотока и др.*). Примером может служить схематическая географическая карта, отдельные районы которой обозначены различным образом в зависимости от *величины* изображаемого статистического признака. *Разновидностью* картограммы является *картодиаграмма*. *Картодиаграмма* – географическая карта, по *отдельным* районам или пунктам которой размещены некоторые графические знаки (*столбики*, *круги и др.*), соответствующие величине статистических показателей, изображенных на ней. К картодиаграммам относятся также *схемы транспортных потоков* различного назначения. Картограммы, на которых размещаются целые таблицы, называются *центрограммами* и позволяют составлять целые статистико-географические описания. *Центрограммы* находят широкое применение при изучении миграции народонаселения и перемещения центров производства различных товаров. Центрографический метод был разработан в начале 20 в. выдающимся русским химиком Д.И. Менделеевым и его сыном И.Д. Менделеевым.

***Диаграмма*** – графическое изображение статистических данных, наглядно показывающее соотношение между сравниваемыми величинами. *Диаграммы* подразделяются по форме и целям представления статистических данных. *Диаграммы сравнения* показывают соотношения различных статистических совокупностей или их частей по какому-либо изменяющемуся в пространстве признаку. Графики данного типа бывают *столбиковые* и *линейные*, размеры компонент которых соответствуют отображаемым ими цифрам. На рис. **9** (**a, b**) представлены примеры обоих видов *диаграмм сравнения* – распределение публикаций **ТТГ** согласно странам в течение 1970 – 1999. ***Структурные диаграммы*** служат для представления структуры *совокупности*. К ним, в первую очередь, относятся *диаграммы удельных весов*, характеризующие отношение отдельных частей совокупности к ее общему объему. По виду они делятся на *столбиковые* и *секторные*. Так, на рис. **9** (**c, d**) представлены примеры *обоих* видов *структурных* диаграмм – распределение публикаций **ТТГ** согласно их типу. Динамические диаграммы отражают изменение явлений во времени и могут быть *столбиковыми*, *линейными* или *спиральными*.

Особое же место среди графических объектов занимают ***контрольно-плановые графики,*** которые из-за их привязки ко времени могут быть отнесены к ***динамическим диаграммам***. На рис. **9** (**e, f**) представлены примеры *динамических диаграмм* – распределение публикаций **ТТГ** по пятилетним периодам. Например, диаграмма рис. **9** (**f**) отлично иллюстрирует динамику публикаций **ТТГ** по пятилеткам по отношению к предполагаемой средней. Последний пример на рис. **9** (**g, h**) иллюстрирует графическое представление процентное *распределения* активности **ТТГ** относительно места издания в течение 1995 - 1999. Наконец, ***графики связи*** будут рассмотрены ниже при изучении вопросов корреляции и регрессии. Графический язык, в общем случае, имеет хорошие выразительные средства и способность к дальнейшему развитию. Поэтому читатель для своих конкретных нужд может и вполне самостоятельно разрабатывать графические объекты, наилучшим образом отвечающие специфическим чертам конкретных задач статистического анализа данных.

Для обеспечения потребностей в графическом представлении статистических данных, современное программное обеспечение (*математические и статистические пакеты для* **ПК,** *прежде всего*) располагает рядом эффективных средств для создания графических объектов наиболее часто используемых типов и для преобразования графических объектов из одного типа в другие. В частности, даже такое средство как *Microsoft Graph*, будучи *Microsoft Office*,

позволяет легко создавать, редактировать и рисовать *диаграммы* наиболее часто используемых типов на основе пользовательских данных различного характера.

Начальные данные для диаграммы могут быть введены как непосредственно с консоли **ПК,** так и из файлов данных, предварительно подготовленных в среде популярных пакетов таких как *Microsoft Excel*, *Lotus 1-2-3* или из файлов данных *ASCII*-типа. Все это существенно облегчает задачу графического представления статистических данных любого характера и назначения. Систематический анализ *диаграмм* как визуального представления фактических данных различного характера может быть найден в книгах [218, 251]. В частности, анализ показывает, что процесс построения графика может быть разделен на три этапа: (**1**) *этап классификации данных*, (**2**) *этап графического проектирования* и (**3**) *этап непосредственного построения графика*. Эти три этапа, выполненные в указанном порядке, более или менее адекватно отражают процесс проектирования графика специалистом. Они также служат основанием для конструктивной теории проектирования диаграммы. Данный вопрос будет рассматриваться более подробно несколько ниже с акцентом на его прикладных аспектах. В частности, современные теоретические и прикладные проблемы анализа и визуализации символьных данных рассматриваются достаточно подробно в книге [295].

Между тем, эффективное использование *графического представления* статистических данных предполагает овладение методикой и техникой их построения. При этом, важно учитывать тот факт, что построение графических изображений является достаточно трудоемким. При этом, следует отметить, что построение графического представления статистических данных, которое в наибольшей степени соответствовало бы характеру и содержанию изображаемых данных и поставленной задаче их анализа, обычно удается не сразу и приходится составлять несколько его вариантов. Существенно ускорить и упростить процесс создания графических представлений статистических данных позволяет применение **ПК,** снабженных *специальными пакетами прикладных программ* (**ППП**). Имеется целый ряд **ППП,** посредством которых можно после ввода исходных данных в **ПК** и выбора конкретного вида графического изображения автоматически получить соответствующее графическое изображение введенных данных. В этом отношении огромные возможности для автоматического построения различных видов графических изображений статистических данных предоставляет программа обработки электронных таблиц *Microsoft Excel* – самая популярная и широко используемая во многих странах мира. К несомненным достоинствам *Microsoft Excel* следует отнести наличие весьма развитой и эффективной справочной системы, которая позволяет пользователю получать необходимую информацию в виде достаточно подробных комментариев, подсказок, что в значительной мере облегчает и ускоряет ее изучение. Это позволяет быстро освоить технику автоматического построения различных видов статистических графиков с помощью пакета *Microsoft Excel* пользователям **ПК,** имеющим различный уровень компьютерной подготовки.

Использование графического представления статистических данных как в статике, так и в их динамике позволяет не только описывать их наиболее понятным и общедоступным языком, но и делать на его базе целый ряд важных предварительных заключений и предположений о характере поведения исследуемых данных. Более того, использование графического метода в целом ряде важных случаев позволяет эффективно получать приближенные с той или иной степенью точности *функциональные зависимости* (*т.н. нами* ***метод компьютерной подгонки***), описывающие статистические данные, которые впоследствии можно успешно использовать или в качестве основы для дальнейшей разработки функциональных связей, либо в качестве приемлемого окончательного решения. На основе этого метода нами для ряда важных задач были получены интересные с практической точки зрения результаты. Несколько ниже будет представлен принцип такого графического метода и некоторые результаты его применения.

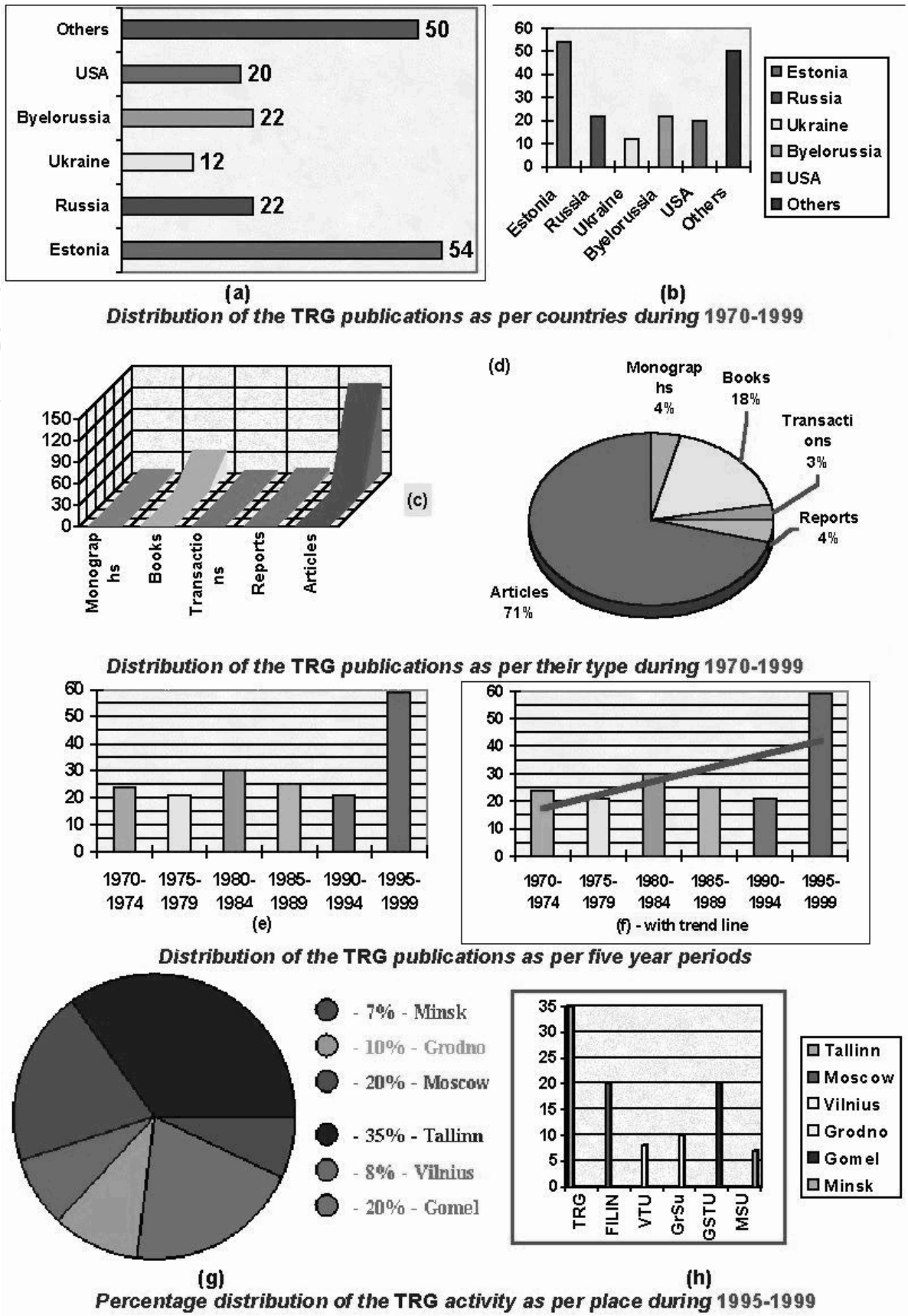



Рис. 9. Примеры различных типов статистических диаграмм

# Глава 5.
# Абсолютные и относительные статистические величины

Количественная определенность явлений выражается как в *абсолютных*, так и *относительных* категориях, называемых ***статистическими величинами***. Так как обширный класс социально-экономических явлений неограничен, характеризуется множеством конкретных типов и форм проявления их абсолютных и относительных величин, то статистика для их изучения выработала широкую систему количественных категорий. В общем случае *статистические величины* (*в дальнейшем просто* ***"величины"***) делятся на два больших класса: ***абсолютные*** и ***относительные***. Любой количественный статистический показатель относится к одному из этих классов. Внутри каждого из указанных классов выделяются подклассы. В ряде случаев отнесение величины к тому или иному классу весьма условно из-за их обладания чертами как одного, так и другого класса. Ниже оба класса величин обсуждаются более детально.

## 5.1. Абсолютные статистические величины

***Абсолютные величины*** – форма количественного выражения статистических показателей, непосредственно характеризующая абсолютные размеры социально-экономических явлений, их признаков в единицах соответствующих систем измерения. Выбор единиц измерения определяется сущностью изучаемого явления и задачами исследования. *Абсолютные величины* – это именованные числа, например, 530 млн. т. нефтепродуктов, 800 млн. кв. м. тканей, 955 тыс. пар обуви, 350 публикаций и др. *Абсолютные величины* составляют основу расчетов различных *относительных величин*, *аналитических* и *обобщающих показателей*. Различают три вида абсолютных величин: *индивидуальные*, *групповые* и *общие*.

***Индивидуальными*** называют *абсолютные величины*, характеризующие размеры *количественных* признаков единиц совокупности. Тогда как ***групповые*** или ***общие*** абсолютные величины характеризуют размеры признака по отношению к группе или всей совокупности в целом, например, итоговые суммы или средние. ***Единицы измерения*** абсолютных величин весьма разнообразны и определяются как объектом, так и задачами исследования. Статистика использует большое число разнообразных единиц измерения, которые тем не менее, в общей классификации *сводимы* к **3-5** типам. Наиболее типичной является разбивка *единиц измерения* на **3** типа, а именно: *натуральные*, *денежные* и *трудовые*. Сюда следует добавить еще два типа: единицы измерения *времени* и *объема* совокупности и ее частей.

***Натуральные*** единицы измерения соответствуют общепринятым физическим системам, например, системе СИ. В рамках натуральных широко используются и ***условные*** единицы измерения. Пересчет в *условные единицы* получил широкое распространение в статистической практике и в научных исследованиях. В статистике используются самые разнообразные пересчеты для получения *условных единиц*. Например, *метрическая тонна* топлива, имеющего

теплотворность (**Т**) в 7000 килокалорий, принимается за единицу, а остальные виды топлива пересчитываются в условные по отношению к ней. Так, метрическая тонна топлива с Т = 9599 килокалорий соответствует примерно (9599/7000)=1.37 условной тонне. При *анализе динамики* научных работ также необходимо вводить условные единицы, чтобы сопоставлять научные публикации различного характера (*монографии*, *книги*, *сборники*. *отчеты*, *статьи и др*.).

***Денежные** единицы* измерения используются для *стоимостной характеристики* статистических показателей. Это позволяет производить соизмерения самых разнородных величин, но изменение цен со временем затрудняет производить сопоставимые оценки. С этой целью в статистике используются сопоставимые цены, получаемые путем некоторой переоценки в цены определенного (*базового*) периода времени. ***Трудовые*** единицы используются для измерения различного рода трудозатрат, а также определения величины трудовых ресурсов, рациональности их использования и т.д. (*человеко-час/день/год и др*.). В ряде случаев единицы измерения и совокупности могут совпадать. Например, единицей человеческой *совокупности* и единицей измерения ее величины является отдельный человек.

## 5.2. Относительные статистические величины

***Относительная*** статистическая величина является мерой количественного соотношения статистических показателей и отражает относительные размеры социально-экономических явлений. Она получается как частное от деления одной величины, называемой обычно *текущей* или *отчетной*, на другую, называемую *базисной*, *базой сравнения* или *основанием*. В качестве текущей и базисной могут выступать как одноименные, так и разноименные величины. В первом случае получаем ***безразмерные*** величины, во втором – ***размерные***. Например, доля части совокупности к ее полному объему есть безразмерная величина или процент, тогда как плотность населения – уже размерная величина (*чел/кв.км*.). При этом, если у основания используется множитель 100 и 1000, то относительная величина выражается соответственно в *процентах* (%) и *промиле* (‰). Могут использоваться для удобства и другие кратные десяти множители.

Относительная *величина планового задания* (**ВПЗ**) есть отношение значения показателя, устанавливаемого на планируемый период, к его величине, достигнутой к этому периоду. Как правило, величина выражается в *процентах*. Строго говоря, относительные величины этого вида не являются статистическими показателями, однако их принято рассматривать в статистике из-за их тесной связи со статистическими *относительными величинами*, например, с показателями выполнения плана. Так, если за 1995 – 1999 годы **ТТГ** подготовила и издала 18 книг и монографий, то относительная величина планового задания в 106% четко говорит о необходимости издания уже [106*18/100] = 19 книг за следующую пятилетку (2000 - 2004). Между прочим, в течение данного периода члены **ТТГ** совсем немного не достигли до запланированного количества 19, издав в США, России, Белоруссии, Эстонии, Литве только 17 книг и монографий [332].

Относительная *величина выполнения плана* (**ВВП**) есть отношение фактического значения показателя к плановому. Данная величина обычно выражается в процентах и сравниваемые значения должны быть сопоставимыми. Например, если планировалось опубликовать за предыдущую пятилетку 15 книг, а было издано 18, то **ВВП** составляет 18*100/15 = 120%**,** т.е. план изданий исследовательской группы *перевыполнен* на 20%**,** если вообще можно говорить о планировании в принятых понятиях творческой активности.

Относительная *величина динамики* (**ВД**) есть отношение значения показателя за данный период времени (*год*, *квартал*, *месяц и т.д*.) к его значению за предыдущий период. Для вычисления этой величины следует иметь данные за два периода времени. При вычислении

**ВД** возникает вопрос о выборе базы: либо она будет *постоянной*, либо *переменной*. В первом случае относительные величины динамики называются *базисными*, во втором – *цепными*. Для иллюстрации сказанного вычислим базисные и цепные **ВД** публикаций **ТТГ** по пятилеткам (рис. **9, e**). В первом случае в качестве *базового* выбираем период 1970 – 1974 годы, во втором - происходит сравнение текущего периода с предыдущим. Получаем *результирующую* табл. **4**.

**Таблица 4. Динамика научных публикаций ТТГ по пятилеткам (1970 – 1999)**

| *Показатель активности исследовательской группы* | **1970-1974** | **1975-1979** | **1980-1984** | **1985-1989** | **1990-1994** | **1995-1999** |
|---|---|---|---|---|---|---|
| *Общее кол-во публикаций* | **24** | **21** | **30** | **25** | **21** | **59** |
| *В % к предыдущему периоду* | **-** | **87.5** | **142.9** | **83.3** | **84** | **281** |
| *В % к базовому периоду (1970 – 1974)* | **100** | **87.5** | **125** | **104.2** | **87.5** | **245.8** |

Теперь, мы можем определить взаимосвязи между **ВПЗ, ВВП** и **ВД,** сделав запись весьма очевидных отношений на основе их определений, а именно:

**ВПЗ = ps/bs, ВВП = fs/ps, ВД = fs/bs, ВД = QEP* ВПЗ**

где **bs** – значение *базисного показателя* в предыдущем периоде, **ps** – *планируемое* значение показателя на данный период и **fs** – его *фактическое* значение в данном периоде. Таким образом, по любым двум данным величинам всегда можно вычислять третью. Например, **ТТГ** планировала на пятилетку 1980 – 1984 годы опубликовать на 10% больше работ по сравнению с первой пятилеткой ее деятельности (1970 – 1974). Фактически же число публикаций увеличилось даже на (125-100)%=25%**.** Тогда **ВД** = 125%**, ВПЗ** = 110% и **ВВП** = (**ВД/ВПЗ**)*100% = 113.6%**,** т.е. фактическое выполнение плана публикаций **ТТГ** составило 113.6% (*если вообще уместно говорить о планах в научном творчестве*).

Относительная *величина структуры* (**ВС**) отпределяет соотношение размеров частей и всей совокупности. Данный показатель называется также *долей* или *удельным весом*. Так, на рис. **9** (**c, d**) можно увидеть удельные веса публикаций **ТТГ** за 1970 – 1999 годы в разрезе их типов: *монографии*, *книги*, *сборники*, *отчеты* и *статьи*. С помощью величины структуры возможно выявлять не только структуру изучаемой совокупности, но и структурные сдвиги в ней, для чего анализируются показатели структуры за несколько периодов времени. Например, анализ динамики по пятилеткам изменения долей публикаций **ТТГ** по их типу позволяет сделать интересный вывод (табл. **5**): *со временем доля статей в периодической печати убывает при соответствующем росте других типов публикаций* (*монографии*, *книги*, *отчеты* и *сборники*). Это вполне согласуется с естественным научным *"взрослением"* группы, отражаемым ростом публикаций обобщающего, монографического характера при относительном уменьшении доли научных работ статейного, более узкого характера.

**Таблица 5. Динамика по пятилеткам доли публикаций ТТГ по их типам**

| *Типы публикаций ТТГ* | **1970-1974** | **1975-1979** | **1980-1984** | **1985-1989** | **1990-1994** | **1995-1999** |
|---|---|---|---|---|---|---|
| *Научные статьи* | **91.6** | **81.0** | **80.0** | **68.0** | **38.1** | **69.5** |
| *Монографии, книги, сборники и др.* | **8.4** | **19.0** | **20.0** | **32.0** | **61.9** | **30.5** |
| *Итого:* | **100** | **100** | **100** | **100** | **100** | **100** |

Представив оба ряда распределения долей для групп публикаций (*статьи* и *другие*) на графической диаграмме наряду с соответствующими тенденциями, приближенными полиномиальными кривыми (рис. **9.1**), мы можем легко визуально вообразить динамику обоих творческих процессов. Из данной диаграммы, в частности, следует, что на интервале 1970 – 1994 имела место вышеупомянутая динамика, которая довольно симметрична относительно уровня в 50%. В то время как на интервале 1995 – 1999 характер динамики изменился с точностью до наоборот. Более того, последнее обстоятельство можно объяснить как существенным увеличением зондирования новых направлений, так и изданием монографий и книг большого объема (400-800 *страниц журнального формата*).

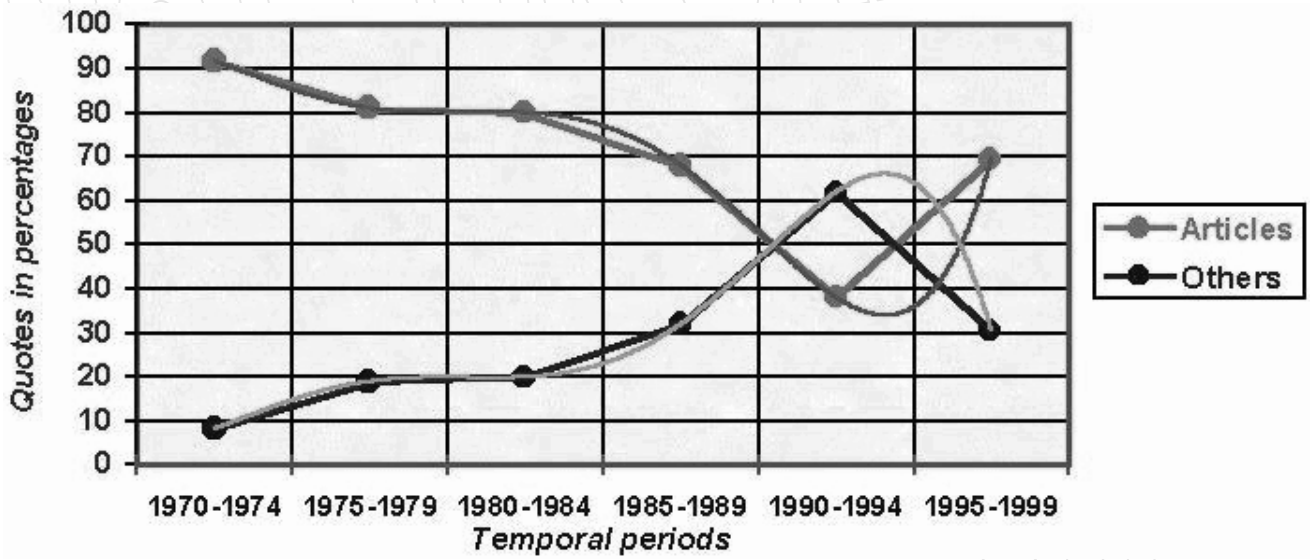



**Рис. 9.1. Динамика публикаций ТТГ по их типам совместно с их трендами**

Относительная *величина координации* (**ВК**) определяет соотношения частей совокупности между собой. Для этого одну из частей совокупности принимают за некую *базу сравнения* и определяют отношение к ней всех других ее частей. В отличие от *величины структуры величины координации* характеризуют не структуру совокупности, а соотношение ее частей между собой, представляя особый вид относительных величин.Так, выбрав в совокупности публикаций **ТТГ** (табл. **1**) за базовую часть *монографии* (**М**), определяем отношение к ней других ее частей: *книг* (**К**), *научных отчетов* (**НО**), *сборников статей* (**СС**) и *статей* (**С**), а именно: **К/М** = 33/7 = 4.7, **НО/М** = 8/7 = 1.1, **СС/М** = 5/7 = 0.7 и **С/М** = 127/7 = 18.1, что позволяет делать уже вполне определенные выводы.

Относительная *величина интенсивности* (**ВИ**) есть степень распространения (*развития*) какого-либо явления в определенной среде, т.е. степень насыщенности определенной среды данным явлением. Величина интенсивности всегда есть отношение разноименных величин – {*величина явления* (*показателя*)/*объем среды*}. Примером данного типа относительных величин может служить коэффициент рождаемости, показывающий, сколько родившихся в течение года приходится на 1000 человек населения.

Относительная *величина сравнения* – отношение одноименных величин, характеризующих весьма разные объекты. Например, сравнение показателей публикаций **ТТГ** с аналогичными показателями других *научных школ*, сравнение численности населения отдельных государств, городов, областей и т.д.

Относительные *величины уровня экономического развития* – показатели, характеризующие размеры производства важнейших видов продукции на душу населения. При этом, иногда

данные показатели относят к относительным *величинам интенсивности*. Между тем, они в определенной мере характеризуют уровень экономического развития страны. Примерами такого типа величин являются производства на душу населения, например: электроэнергии, товаров народного потребления, продуктов питания и др.

*Относительные* величины – один из важнейших способов *обобщения* и *анализа* статистических данных, выбор их вида определяется целями и объектом исследования. При их выборе следует соблюдать ряд основных требований [64], а корректность применения требует рассмотрения их во взаимосвязи друг с другом. При этом, особую роль играет рассмотрение относительных величин в совокупности с определенными объемными величинами.

Таким образом, ***абсолютные статистические величины*** выражают размеры (*уровни*, *объёмы*) явлений и процессов. Любая статистическая информация начинает формироваться именно с абсолютных величин. В зависимости от целей анализа величины могут быть натуральные, денежные (*стоимостные*), трудовые. Между тем, изучая социально-экономические явления, нельзя ограничиться только *абсолютными величинами*. В анализе статистической информации важное место занимают ***относительные величины.*** Относительные величины представляют собой ***частное*** от деления двух статистических величин, характеризующее количественное отношение между ними, в числителе всегда находится показатель, который изучается, а в знаменателе с чем сравнивается, этот показатель называется *базой сравнения*. Относительная величина (*результат сравнения*) может быть выражена в процентах или в виде коэффициента. При расчёте относительных показателей весьма важно обеспечить сопоставимость числителя и знаменателя. *Относительные показатели* находят большое применение, а именно они могут использоваться: (**1**) при изучении структуры статистической совокупности, (**2**) для изучения динамики (*т.е. для характеристики изменения изучаемого явления во времени*), (**3**) для сравнения одноименных показателей, но относящихся к различным объектам статистического анализа, (**4**) для характеристики интенсивности изучаемого явления (*т.е. насколько широко исследуемое явление распространено в той или иной среде*). При работе с абсолютными и относительными величинами важным условием является комплексное их применение, если рассматривать их порознь, то можно прийти к неправильным выводам, и только их совместное применение даёт всестороннюю, объективную характеристику изучаемого объекта либо явления.

# Глава 6.
# Основы метода средних величин

***Средние величины*** играют исключительно важную роль в статистике. ***Метод средних*** в его общей форме, как и метод *группировок*, является *специфической* особенностью статистической методологии. ***Средняя*** величина представляет собой обобщенную характеристику *признака* в статистической совокупности, единицы которой подвержены действию *различных* факторов. ***Средняя*** является общей мерой их действия, их равнодействующей. В *средней* величине массового явления *нивелируются индивидуальные* различия единиц совокупности в значениях осредняемого признака, поэтому в ней проявляются общие закономерности, присущие данной совокупности.

***Средняя*** является важнейшей категорией статистики и важнейшей формой обобщающих показателей. Средняя выступает важнейшим методом обобщения и в этом смысле говорят о ***методе средних величин,*** широко применяемом в экономико-статистических исследованиях. Математические приемы, используемые в различных разделах статистики, непосредственно связаны со *средними величинами*. Со *средними величинами* тесно связаны многие аналитические исследования в статистике и связанных с нею областях анализа различных характера и типа данных. В настоящей главе мы сконцентрируем наше внимание на *методе средних* в целом.

## 6.1. Свойства средней арифметической

***Средняя арифметическая*** – наиболее распространенный в статистике тип средних величин. Как и все другие средние величины, она применяется в форме *простой* и *взвешенной* средней (*средневзвешенной*). Положим, что конечное множество **A = {a1, a2, ..., an}** есть набор вариантов некоторого *признака* явления (*при этом, среди элементов множества* **A** *могут быть идентичные*). Тогда *средняя арифметическая* ($\bar{A}$) и *средневзвешенная арифметическая средняя* ($\tilde{A}$) вычисляются соответственно по следующим простым формулам:

$$\bar{A} = \frac{\sum_{j=1}^{n} a_j}{n} \qquad \tilde{A} = \frac{\sum_{k=1}^{m} a_k f_k}{\sum_{k=1}^{m} f_k} \tag{31}$$

где $f_k$ – *частота* (*вес*) варианта $a_k$ признака ($k = 1 .. m$). Сразу же сделаем уточнение, если простая средняя независимо от ее типа используется для первичных рядов распределения, то средневзвешенная – для *вариационных рядов* (**ВР**). Пусть теперь, не нарушая общности, **X**-множество вариантов разбито на два подмножества $B = \{b_1, b_2, ..., b_m\}$ и $C = \{c_{(m+1)}, c_{(m+2)}, ..., c_n\}$ таких, что для них справедливы следующие соотношения: $A = B \cup C$ и $B \cap C = \emptyset$ ($\emptyset$ – пустое множество). Очевидно:

$$\bar{B} = \frac{\sum_{j=1}^{m} b_j}{m} = \frac{S1}{m} \qquad \bar{C} = \frac{\sum_{k=m+1}^{n} C_k}{n-m} = \frac{S2}{n-m} \qquad \frac{S1+S2}{n} = \bar{A} \tag{32}$$

Тогда на основе формул (31, 32) легко вычисляем среднюю средних данных подмножеств **B** и **C,** а именно:

$$\overline{A}' = \frac{\overline{B} + \overline{C}}{2} = \frac{(n-m)S1 + mS2}{2m(n-m)} \qquad (33)$$

Не нарушая общности, полагаем, что $m = min\{n - m, m\}$, и, проведя достаточно несложные преобразования, получаем следующее соотношение:

$$\overline{A}' = \frac{n\overline{A}}{2(n-m)} = \frac{(n-2m)S1}{2m(n-m)}$$

Откуда легко получаем, что равенство $\bar{A} = \bar{A}`$ имеет место тогда и только тогда, когда $m = n/2$. В общем случае формулировка принимает следующий вид: *если совокупность* **A** *разбивается на* **m** *непересекающихся групп* $A_k$ *по* $N_k$ *единиц в каждой, т.е.* $\sum_k N_k = m$, *то величины* $\bar{A}$ и $\bar{A}` = \sum_k A_k / m$ *совпадают тогда и только тогда, когда* $(\forall k \in \{1, 2, \ldots, n\})(N_k = n/m)$, *исключая тривиальный случай отсутствия вариации у признака*. В нашей трактовке *первую* формулу (31) вполне можно рассматривать как *базовую среднюю* всей **A**-совокупности, а вторую – как *среднюю*, *устраняющую* различия между *базовой средней* и *усредненной групповой*.

По табл. **1** (*графа* **4**) подсчитаем *усредненную групповую* затрат на подготовку публикаций **ТТГ** (11.25+13.6+17.5+38.5+35.4)/5 = **23.25,** тогда как для средней взвешенной (*точнее базовой средней совокупности*) получаем значение **24.4**. В данном примере разница обоих типов средней не столь велика, но в случае больших совокупностей, значительных диапазонов вариации и неравномерных группировках это может приводить к весьма существенным различиям. Единственное указанное нами исключение из правила является достаточно редким, а в статистике и вовсе исключается из рассмотрения. Рассмотрим основные свойства *средней* и *средней взвешенных арифметических*, имеющие чрезвычайно большое практическое значение.

**Свойство 1.** *При изменении всех вариантов величины* **A** *в* **m** *раз* (*на число* **m**), *ее средняя* (*средневзвешенная*) $\tilde{A}$ *также изменяется в* **m** *раз* (*на число* **m**).

**Свойство 2.** *При изменении в* **m** *раз весов всех вариантов величины* **A** *не изменяет значения* $\tilde{A}$ *ее средневзвешенной*

**Свойство 3.** *Для вычисления средневзвешенной* $\tilde{A}$ *вместо частот* $f_k$ *вариантов можно использовать их частости* (*доли*) $p_k = f_k / \sum_k f_k$.

**Свойство 4.** *Средневзвешенная* $\tilde{A}$, *умноженная на численность* **A**-*совокупности*, *равняется сумме произведений каждого варианта на его частоту*.

**Свойство 5**. *Сумма отклонений индивидуальных значений признака* **A** *от их средней* $\bar{A}$ (*средневзвешенной* $\tilde{A}$) *равна нулю*.

Доказательство свойств **1 - 5** средней вытекает из *определения* самой средней (*средневзвешенной*) арифметической и предлагается читателю в качестве полезного несложного упражнения {*см. формулы* (31)}.

**Свойство 6.** Пусть $A = \{a1, a2, \ldots, an\}$. Величина $L = \sum_k (a_k - G)^2$ минимальна при значении $G = \bar{A}$. Доказательство свойства **6** состоит в нахождении минимума функции $L = L(G)$, а именно:

$$\frac{\partial L(G)}{\partial G} = -2\sum_{k=1}^{n}(a_k - G) \qquad \sum_{k=1}^{n} a_k - G*n = 0 \qquad G = \frac{\sum_{k=1}^{n} a_k}{n} = \overline{A}$$

Свойства **1 - 4** позволяют в ряде случаев упрощать вычисления средних. Так, свойства **1** и **2** позволяют упрощать вычисления средних в случае больших значений или частот вариантов признака. Свойство **3** позволяет вычислять средние при неизвестных абсолютных значениях весов, но известных между ними пропорциях или других связанных с ними значениях. Свойство **4** широко используется на практике при расчетах, в планировании и т.п. Например, средняя зарплата, умноженная на число работников, дает фонд зарплаты; средняя урожайность, умноженная на посевную площадь, дает валовой сбор и т.п. В случае разбиения значений признака по *интервалам* расчет *средневзвешенных* производится на основе их *среднеинтервальных* значений. Данное замечание распространяется на все последующее изложение.

*Средняя* (*средневзвешенная*) *арифметическая*, с вычислением которой мы познакомились выше, является наиболее простой и в то же время наиболее всеобщей формой средней, но далеко не единственным видом *средних*. Практика и теория статистики показывают, что применение во всех без исключения случаях этого типа *средних* приводит к очень грубым ошибкам. Поэтому во многих случаях требуется другая методика вычисления средних величин.

## 6.2. Другие типы средних величин и их выбор

В зависимости от характера осредняемого признака и имеющихся данных применяются виды средних, отличные от *арифметической* и *взвешенной арифметической*. В статистике наряду с отмеченными применяются преимущественно следующие средние: *гармоническая*, *геометрическая* и *квадратическая*, а также *мода* и *медиана*. Как и прежде, рассматривается статистическая совокупность $\mathbf{A = \{a1, a2, ..., an\}}$; при этом, частоты $\mathbf{f_k}$ вариантов признака **A** могут отличаться от единицы.

*Средняя* (*средневзвешенная*) ***гармоническая*** вычисляется по следующим простым формулам:

$$\overline{A}_h = \frac{n}{\sum_{j=1}^{n} 1/a_j} \qquad \tilde{A}_h = \frac{\sum_{k=1}^{m} f_k}{\sum_{k=1}^{m} f_k/a_k} \tag{34}$$

*Средняя гармоническая* применяется тогда, когда необходимые веса в исходных данных явно не заданы, а входят сомножителем в один из имеющихся показателей. Например, в трех фирмах общий фонд (**WF**) зарплаты и среднемесячная зарплата (**AW**) работающего в EEK (*эстонские кроны*; 1 USD ≈ 16 EEK на 16.03.2000**;** на 23.01.2006 курс составлял 1 USD ≈ 12.76 EEK) соответственно равны: **WF1** = 230.000, **WF2** = 260.000, **WF3** = 250.000 и **AW1** = 2400, **AW2** = 2000, **AW3** = 2440. Требуется вычислить *среднюю* зарплату (**AW**) работников этих трех фирм. Применяя среднюю *арифметическую*, получаем **AW** = (2400 + 2000 + 2440)/3 = 2280. Но этот результат не адекватен искомому, определяемому из общей расчетной формулы

**AW = (*общий фонд зарплаты*)/(*общая численность работников*)**

на основе которого получаем следующее простое соотношение:

$$\mathbf{AWh} = \frac{\mathbf{WF1+WF2+WF3}}{\frac{\mathbf{WF1}}{\mathbf{AW1}}+\frac{\mathbf{WF2}}{\mathbf{AW2}}+\frac{\mathbf{WF3}}{\mathbf{AW3}}}$$

Это и есть формула *средневзвешенной гармонической* и, сделав необходимые вычисления по ней, получаем искомое значение **AWh** = 2254.1, которое весьма существенно отличается от первоначального результата.

В связи с приведенными колебаниями курса доллара относительно ЕЕК приведем одно существенное соображение. До настоящего времени США фактически были мировым монетным двором, а доллар был признанным мировым средством оплаты. Финансовая и экономическая система США базировались на стабильности экономики и твердых позициях доллара. Однако, с конца 2003 начинается заметный спад экономики США и долгосрочная нестабильность доллара с очевидной тенденцией к понижению его курса относительно ведущих валют мира. Это наряду с появлением евро, растущая неприязнь к США из-за ее высокомерия и амбициозных претензиях на мирового судью, и т.д. весьма реально может привести к отказу от доллара, как основного *мирового платежного средства*, что в совокупности с другими отрицательные факторы для США *несомненно* могут привести к *самому* серьезному кризису всей *финансовой* и *экономической* системы США в перспективе совсем не отдаленного будущего.

*Гармоническая средняя* часто применяется для определения средних скоростей. Например, первые 30 км автомобиль проезжает со средней скоростью 120 км/ч, а остальные 70 км – со средней скоростью 110 км/ч. Для вычисления в таком случае средней скорости используем средневзвешенную гармоническую величину **Vh** = (30 + 70)/(30/120 + 70/110) = 112.8 км/ч. В практике советской статистики по гармонической средней вычислялись средневзвешенные индексы государственных цен. Применение данного типа средней оправдано при расчетах средних: трудоемкости единицы продукции, продолжительности строительства объектов и ряда других важных статистических показателей.

***Квадратическая средняя*** (*средневзвешенная*) вычисляются по следующим формулам:

$$\overline{A}_q = \sqrt{\frac{\sum_{j=1}^{n} a_j^2}{n}} \qquad \widetilde{A}_q = \sqrt{\frac{\sum_{k=1}^{m} a_k^2 f_k}{\sum_{k=1}^{m} f_k}} \tag{35}$$

Средняя данного типа применяется для осреднения величин, входящих в совокупность в виде квадратных показателей (*например, при расчетах средних диаметров труб, шлангов, проката, стволов деревьев и др.*). Например, для вычисления среднего диаметра труб (*в см.*) диаметрами 42, 47, 67, 62, 58, 53, 33 и 38 используем *квадратическую среднюю* {*формула* (35)}:

$$\overline{U}_q = \sqrt{\frac{42^2+47^2+67^2+62^2+58^2+53^2+33^2+38^2}{8}} = 51.25$$

Тогда как использование *арифметической средней* дало бы существенно отличную величину $\bar{U}$ = 50 см. *Квадратическая средняя величина* находит весьма широкое применение, в частности, для оценки варьирования признаков в дисперсионном анализе.

***Геометрическая средняя*** (*средневзвешенная*) вычисляется по следующим простым формулам:

$$\ln(\overline{A}_g) = \frac{\sum_{j=1}^{n} \ln(a_j)}{n}; \quad \mathrm{Ln}(\widetilde{A}_g) = \frac{\sum_{k=1}^{m} f_k \ln(a_k)}{\sum_{k=1}^{m} f_k}; \quad \overline{A}_g = \sqrt[n]{\prod_{j=1}^{n} a_j} \tag{36}$$

Для удобства представления формулы приведены в логарифмическом виде, который весьма удобен при практическом вычислении *среднегеометрических* в случае большого числа значений признака. В частности последняя формула (36) представляет традиционную форму *геометрического среднего*. Данный тип средней используется, в частности, в статистике при вычислениях средних темпов роста и при построении индексов. Так, используя табл. **4** и

формул (31, 36), определим с помощью средней *геометрической средний коэффициент* (**Rz**) *роста объема* научных публикаций **ТТГ** (**Z**) по пятилеткам ее творческой активности:

$$\mathbf{Rz} = \frac{\overline{\mathbf{Z}}_{\mathbf{g}}}{\overline{\mathbf{Z}}} = \frac{\sqrt[6]{\frac{24*21*30*25*21*59}{6}} = 27.87}{\frac{24+21+30+25+21+59}{6} = 30} = 0.929$$

Таким образом, практически **Rz** = 0.93, что говорит о достаточно равномерной динамике публикаций **ТТГ** по пятилеткам (*даже несмотря на резкое повышение деятельности в прошлый период*), что, на наш взгляд, весьма неплохой творческий показатель для научной активности исследовательского коллектива.

Для характеристики некоторых совокупностей подходит еще один *своеобразный* тип средней ***– антигармоническая*** (*взвешенная*), вычисляемая по следующим формулам:

$$\overline{\mathbf{A}}_{\mathbf{ah}} = \frac{\sum_{j=1}^{n} a_j^2}{\sum_{j=1}^{n} a_j};\qquad \widetilde{\mathbf{A}}_{\mathbf{ah}} = \frac{\sum_{k=1}^{m} a_k^2 f_k}{\sum_{k=1}^{m} a_k f_k};\qquad \overline{\mathbf{A}}_{\mathbf{ah}} = \frac{\overline{\mathbf{A}}_{\mathbf{q}}^2}{\overline{\mathbf{A}}};\qquad \widetilde{\mathbf{A}}_{\mathbf{ah}} = \frac{\widetilde{\mathbf{A}}_{\mathbf{q}}^2}{\widetilde{\mathbf{A}}} \tag{37}$$

При этом, две последних формулы (37) устанавливают отношения между *антигармонической средней* (*антигармонической взвешенной*), *квадратичной средней* (*квадратичной взвешенной*) и *арифметической средней* число (*арифметической взвешенной*) {*см. формулы* (31, 35)}.

Рассмотрим использование *антигармонической средней* на следующем простом примере. Пусть имеется **n** отраслей и **Kj** – эффективность вложений в **j**-ю отрасль (*т.е. вложенный в текущем году* 1 *рубль дает в следующем году доход в* **Kj** *рублей*). При эьом, если показатели **Kj** постоянны, то *эффективность вложений* (**E**) выражается *антигармонической* средней, вычисляемой по следующей простой формуле:

$$\overline{\mathbf{E}}_{\mathbf{ah}} = \frac{\sum_{j=1}^{n} K_j^2}{\sum_{j=1}^{n} K_j}$$

Предположим, что имеются четыре отрасли с эффективностями капитальных вложений соответственно **Kj** = 1 + 0.18***j** (**j** = 1 .. 5). Тогда на основе полученной формулы мы имеем:

$$\overline{\mathbf{E}}_{\mathbf{ah}} = \frac{\sum_{j=1}^{5} (1+0.18*j)^2}{\sum_{j=1}^{5} (1+0.18*j)} = 1.5821$$

Следует отметить, что возможны дальнейшие расширения понятия средней, позволяющие решать более специфические экономико-статистические задачи. Более детально с данной проблемой заинтересованный читатель ознакомиться в книгах [55, 92, 116, 126, 128, 133, 150, 155, 157, 182]. Между тем, относительно одной и той же **A**-совокупности рассмотренные нами типы *средних* удовлетворяют следующему *правилу мажорантности*:

$$\overline{\mathbf{A}}_{\mathbf{h}} \le \overline{\mathbf{A}}_{\mathbf{g}} \le \overline{\mathbf{A}} \le \overline{\mathbf{A}}_{\mathbf{q}} \le \overline{\mathbf{A}}_{\mathbf{ah}}$$

где знак равенства "**=**"имеет место только в случае отсутствия вариации признака, т.е.:

$$(\forall \mathbf{j})(\mathbf{a_j} = \mathbf{a} \rightarrow \overline{\mathbf{A}}_{\mathbf{h}} = \overline{\mathbf{A}}_{\mathbf{g}} = \overline{\mathbf{A}} = \overline{\mathbf{A}}_{\mathbf{q}} = \overline{\mathbf{A}}_{\mathbf{ah}} = \mathbf{a})$$

Доказательство данного правила можно получить, например, средствами математического анализа, исходя из вида общей формулы *степенных средних* (38). Искушенному читателю это рекомендуется проделать в качестве весьма полезного упражнения. Разница между *средними* растет с ростом вариации осредняемых величин. При небольшой вариации данная разница незначительна, что в ряде случаев позволяет использовать более простые типы средней. Так, при анализе *темпов роста* использование *средней* $\bar{\mathbf{A}}$ более распространено, чем *геометрической* средней $\bar{\mathbf{A}}\mathbf{g}$.

Из-за различных типов средних возникает необходимость их адекватного выбора в каждом конкретном случае. С этой целью вводятся два понятия: *определяющая функция* и *уравнение средней*. ***Определяющая функция*** (**DF**) – *обобщающий показатель* (**ОП**) **A**-совокупности, от которого зависит величина средней ($\bar{\mathbf{A}}$), т.е. $\bar{\mathbf{A}} = \mathbf{DF}(\bar{\mathbf{A}})$. *Определяющая функция* и *средняя* тесно связаны – значение **ОП** не меняется при подстановке в соответствующее уравнение вместо *индивидуальных* значений их *средней*. Выбор *типа* средней состоит из следующих четырех этапов, а именно:

1. *Формулировка задачи, для решения которой вычисляется средняя, и определяющей функции –* **ОП;** *при наличии для совокупности нескольких* **ОП** *можно вычислять для нее несколько типов средних*
2. *Нахождение математического выражения* (*определяющей функция*) *для* **ОП**
3. *Составление уравнения средней путем замены в* **ОП** *индивидуальных значений средними величинами*
4. *Решение уравнения средней и определение конкретной формулы для средней; входящие в уравнение средней величины должны быть связаны по смыслу так, чтобы получилась размерность обобщающего показателя*.

Описанную процедуру *выбора* типа средней рассмотрим на простом примере. Предположим, что имеется следующий простой временной (*динамический*) ряд:

| ***Год акттивности ТТГ (1970 - 1999)*** | ***Количество публикаций*** | ***Темпы роста публикаций*** |
|---|---|---|
| **1970** | $\mathbf{A_1}$ | $\mathbf{A_1}$ |
| **1971** | $\mathbf{A_2}$ | $\mathbf{K_1 = A_2/A_1}$ |
| **1972** | $\mathbf{A_3}$ | $\mathbf{K_2 = A_3/A_2}$ |
| **……** | **……** | **……** |
| **1999** | $\mathbf{A_n}$ | $\mathbf{K_{n-1} = X_n/X_{n-1}}$ |

который четко отражает динамику научных публикаций **ТТГ** по годам ее активности. Для характеристики темпов роста числа публикаций мы используем *коэффициенты роста* (**K**) – *отношения текущего года к предыдущему*, т.е. $\mathbf{K_j = A_{j+1}/A_j}$ for $\mathbf{j \in \{1, 2, \ldots, n-1\}}$. Нам требуется определить среднюю величину роста, т.е. *тип* средней для **K**-коэффициента. Согласно вышесказанному определяем *обобщающий показатель* следующим образом:

$$\mathbf{A_n = A_1 \prod_{j=1}^{n-1} K_j}$$

но тогда после простых преобразований получаем следующие соотношения:

$$A_n = A_1 \prod_{j=1}^{n-1} \overline{K}; \qquad (n-1)\ln(\overline{K}) = \ln(\prod_{j=1}^{n-1} K_j); \qquad Ln(\overline{K}) = \frac{\sum_{j=1}^{n-1} \ln(K_j)}{n-1}$$

Следовательно, для такого класса задач используется *геометрическая средняя*, а не, например, арифметическая средняя. Для упрощения вычисления геометрической средней ($\bar{A}g$) можно применять логарифмирование по формулам (36), заменяя извлечение корня **n**-й степени из произведения *средней арифметической* логарифмов сомножителей с *последующим* возведением в степень полученного результата.

Следует отметить, что формулы для *средних* величин могут быть получены и на основе *общих степенных средних*, определяемых следующими выражениями:

$$\overline{A}_k = \sqrt[k]{\frac{\sum_{j=1}^{n} a_j^k}{n}} \qquad \widetilde{A}_k = \sqrt[k]{\frac{\sum_{p=1}^{m} a_p^k f_p}{\sum_{p=1}^{m} f_p}} \tag{38}$$

где $f_k$ – частоты варианта **A**-признака. В зависимости от значения **k**-показателя мы получаем различные типы рассмотренных средних согласно следующей табл. **6.**

**Таблица 6. Типы средних значений (*величин*)**

| **k**-*значение* | *Определяющая функция* | *Тип средней* |
|---|---|---|
| **-1** | **DF = 1/A** | *гармоническая* |
| **0** | **DF = ln(A)** | *геометрическая* |
| **1** | **DF = A** | *арифметическая* |
| **2** | **DF = $A^2$** | *квадратическая* |

В статистике используются и другие *типы* средних, но значительно реже. Более подробное рассмотрение данного вопроса выходит за рамки настоящей книги, однако *заинтересованный* читатель более детально с данным вопросом может ознакомиться в книгах [55, 92, 116, 126, 128, 133, 150, 155, 157, 182]. Между тем, при использовании *метода средних* следует учитывать ряд их свойств, без чего возможны недостоверные выводы по исследуемой совокупности, ибо средняя не только выявляет общие тенденции, но многое также и нивелирует (*скрывает от первоначального впечатления характер явления или процесса*).

## 6.3. Структурные средние величины совокупностей

Наряду с рассмотренными выше в статистике используются еще две особые разновидности *средних* величин, которые вытекают из характеристики статистических рядов и не являются результатом алгебраических вычислений. Условно их можно назвать *структурными средними* – это *мода* и *медиана*. Структурный характер этого типа средних будет легко усматриваться из дальнейшего изложения. ***Модой*** (**Mo**) называется вариант признака, имеющий наибольшую частоту (*частость*), т.е. *мода – наиболее типичное* значение признака. ***Медиана*** (**Me**) – значение *вариационного ряда* (**ВР**), расположенное в его середине, т.е. медиана делит ряд на две равные части. В отличие от других типов средних, значения которых не обязательно принадлежат совокупности, **Mo** всегда является единицей совокупности – дискретного **ВР**.

*Структурные средние* – особый вид средних величин– применяются для изучения *внутреннего* строения рядов распределения значений признака, а также для оценки средней величины (*степенного типа*), если по имеющимся в наличии статистическим данным ее расчет не может быть выполнен. В качестве структурных средних чаще всего используют показатели ***моды*** – *наиболее* часто повторяющегося значения признака, и ***медианы*** – величины признака, которая делит упорядоченную последовательность его значений на две равные по численности части. В итоге у одной половины единиц совокупности значение признака не превышает *медианного* уровня, а у другой – не меньше его. Если изучаемый признак имеет дискретные значения, то особых затруднений при расчете *моды* и *медианы* нет. Если же данные о значениях признака **X** представлены в виде упорядоченных интервалов его изменения (*интервальных рядов*), расчет моды и медианы несколько усложняется и используется интерполяционный подход.

Наиболее простыми являются вычисления величин **Mo** и **Me** для дискретного **ВР** (*раздел* **4.5**). Для него **Mo** определяется как вариант, имеющий частоту (*частость*), большую, чем для его соседей *слева* и *справа*. В *интервальных* **ВР** для вычисления *величины* **Mo** сначала определяется *модальный интервал* (**МИ**): *если ряды* с *равными интервалами – по наибольшей частоте* (*частости*), при *неравных* – по *наибольшей плотности* распределения. При равных интервалах величина **Mo** внутри **МИ** вычисляется по следующей простой формуле:

$$\mathbf{M_o = A_d + \frac{d*(f-f1)}{2*f-f1-f2}} \qquad (39)$$

где **Ad** – нижняя граница **МИ; f, f1, f2** – частоты (*частости*) соответственно *модального*, *предмодального* и *постмодального* интервалов, **d** – величина интервала **ВР**. *Унимодальный* **ВР** имеет одну **Mo,** *мультимодальный* – несколько. Полноценное смысловое значение величина **Mo** имеет только для *унимодальных* **ВР,** для **V-, U**-образных и *мультимодальных* распределений величина **Mo** вообще не определяется.

Для дискретного **ВР** с (**2n + 1**) членами величина **Me** совпадает с (**n + 1**)-м его членом, а с **2n** членами – равна средне арифметическому значений двух центральных его членов. Перед вычислением величины **Me ВР** должен быть обязательно *ранжирован* в порядке возрастания/ убывания, если до этого он не ранжировался. Для вычисления величины **Me** в *интервальном* **ВР** применяется следующая простая формула:

$$\mathbf{M_e = A_d + \frac{d*(S-2*S1)}{2*f}} \qquad (40)$$

где **Ad** – нижняя граница медианного интервала (*интервал*, *содержащий медиану*), **d** – значение интервала **ВР, S** – общее количество наблюдений, **S1** – общая сумма совокупных частот всех интервалов, предшествующих медианному, **f** – частота медианного интервала. Формулы (39) - (40) выводятся из пропорций, получающихся из простых геометрических соображений при рассмотрении гистограммы *интервального* **ВР**. Следует отметить, что величина **Me** не несет особого смыслового значения, кроме того, что делит **ВР** на две равные части. Иногда также используется *наиболее инерционное значение* (**НИЗ**) – такое значение варианта **ВР,** для которого произведение его значения на частоту является *максимальным* (*момент инерции*). Для случая рассматриваемого ниже **ВР** величина **НИЗ=3.** Для дальнейшего изложения нам понадобится один многоаспектный пример результатов наблюдения – сводка научных публикаций **ТТГ** за 1970 – 1999 годы (табл. **7**).

В Табл. **7,** построенной на основе Табл. **1,** сделана перегруппировка первичных данных [23, 191], а именно, научные публикации представлены в разрезе двух следующих *основных* типов – (**1**) *монографические публикации* (*монографии*, *книги*, *отчеты* и *сборники*) и (**2**) *периодические публикации*, разделенные, в свою очередь, на *отечественные* (**А**) и *зарубежные* (**В**) публикации.

Это было сделано с целью более адекватного анализа динамики публикаций **ТТГ**.

**Таблица 7. Распределение публикаций ТТГ по годам ее активности (1970 – 1999)**

| *Год издания* | *Монографические публикации* | *Объем в страницах* | *Периодические публикации* | | | | *Общее количество публикаций* | *Общий объем публикаций* |
|---|---|---|---|---|---|---|---|---|
| | | | *отечественные (A)* | | *зарубежные (B)* | | | |
| | | | *Количество* | *Объем в страницах* | *Количество* | *Объем в страницах* | | |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| 1970 | 0 | 0 | 3 | 23 | 0 | 0 | 3 | 23 |
| 1971 | 1 | 120 | 8 | 94 | 0 | 0 | 9 | 214 |
| 1972 | 1 | 235 | 3 | 18 | 0 | 0 | 4 | 253 |
| 1973 | 0 | 0 | 2 | 14 | 0 | 0 | 2 | 14 |
| 1974 | 0 | 0 | 3 | 29 | 3 | 49 | 6 | 78 |
| 1975 | 1 | 180 | 2 | 12 | 1 | 22 | 4 | 214 |
| 1976 | 1 | 310 | 4 | 93 | 1 | 21 | 6 | 424 |
| 1977 | 0 | 0 | 1 | 8 | 0 | 0 | 1 | 8 |
| 1978 | 1 | 100 | 6 | 83 | 1 | 34 | 8 | 217 |
| 1979 | 0 | 0 | 0 | 00 | 1 | 22 | 1 | 22 |
| 1980 | 2 | 376 | 4 | 75 | 0 | 0 | 6 | 451 |
| 1981 | 1 | 160 | 6 | 135 | 2 | 27 | 9 | 322 |
| 1982 | 1 | 285 | 2 | 7 | 2 | 17 | 5 | 309 |
| 1983 | 2 | 220 | 5 | 234 | 1 | 6 | 8 | 460 |
| 1984 | 0 | 0 | 0 | 0 | 3 | 30 | 3 | 30 |
| 1985 | 3 | 565 | 0 | 0 | 1 | 8 | 4 | 573 |
| 1986 | 2 | 320 | 0 | 0 | 3 | 58 | 5 | 378 |
| 1987 | 1 | 50 | 0 | 0 | 2 | 45 | 3 | 95 |
| 1988 | 3 | 312 | 2 | 25 | 8 | 132 | 13 | 469 |
| 1989 | 2 | 602 | 0 | 0 | 1 | 11 | 3 | 613 |
| 1990 | 2 | 762 | 0 | 0 | 1 | 23 | 3 | 785 |
| 1991 | 3 | 702 | 0 | 0 | 1 | 7 | 4 | 709 |
| 1992 | 3 | 561 | 0 | 0 | 0 | 0 | 3 | 561 |
| 1993 | 5 | 1156 | 2 | 14 | 0 | 0 | 7 | 1170 |
| 1994 | 1 | 115 | 0 | 0 | 0 | 0 | 1 | 115 |
| 1995 | 2 | 450 | 4 | 62 | 2 | 23 | 8 | 535 |
| 1996 | 2 | 558 | 5 | 53 | 3 | 18 | 10 | 629 |
| 1997 | 4 | 1192 | 5 | 47 | 3 | 12 | 12 | 1251 |
| 1998 | 4 | 1196 | 6 | 52 | 4 | 18 | 14 | 1266 |
| 1999 | 5 | 3275 | 6 | 58 | 4 | 21 | 15 | 3354 |
| **Итого** | **53** | **13802** | **79** | **1136** | **48** | **604** | **180** | **15542** |

В качестве простого примера рассмотрим вычисление величин **Mo** и **Me** для интервального **ВР A,** составленного на основе **2**-й графы табл. **7,** а именно:

| *интервал:* | 0 - 1 | 1 - 2 | 2 - 3 | 3 - 4 | 4 - 5 | 5 - 6 |
|---|---|---|---|---|---|---|
| *частота:* | 6 | 9 | 7 | 4 | 2 | 2 |

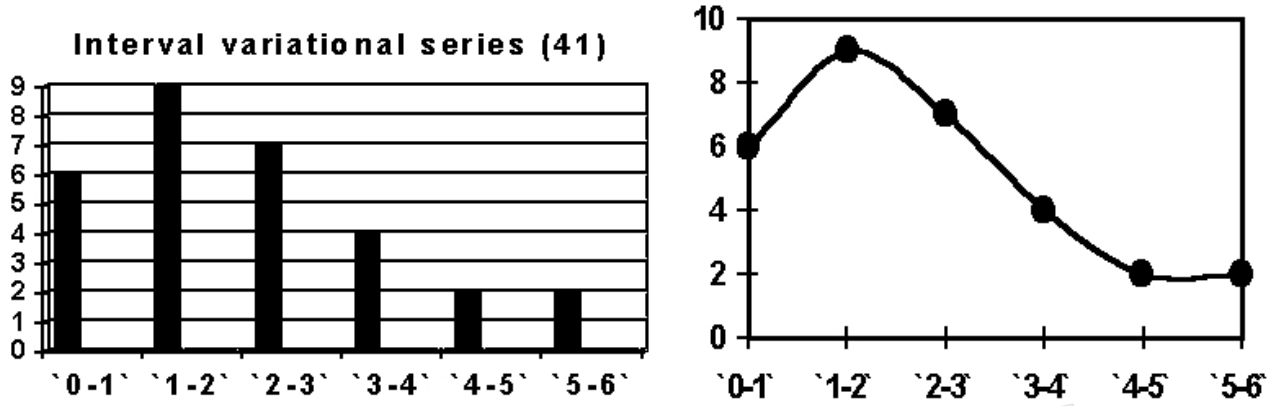
**(41)**

Нетрудно убедиться, что *модальный* и *медианный* интервалы совпадают с интервалом "**1 - 2**" данного **ВР** (**41**). Используя теперь формулы (39) и (40), получаем значения соответственно для величин **Mo, Me** и **Ā,** а именно:

$$Mo = 1 + \frac{1*(9-6)}{2*9-6-7} = 1.6; \qquad Me = 1 + \frac{1*(30-2*6)}{2*9} = 2$$

$$\bar{A} = \frac{0*6+1*9+2*7+3*4+4*2+5*2}{6+9+7+4+2+2} = 1.77$$

Из полученных результатов видно, что значения величин **Mo, Me** и **Ā ВР** (41) весьма близки между собой, что позволяет, *вроде бы*, говорить о том, что соответствующее ему *распределение* довольно близко к нормальному, для которого имеет место соотношение **Mo = Me = Ā**. Тогда как в разделе **7.1** будет показано, что это далеко не так.




В целом ряде случаев величины **Mo** и **Me** являются более адекватными характеристиками **ВР**. Например, при определении объема производства и реализации наиболее ходовых по размерам товаров (*обувь*, *одежда и др*.) показатель **Mo** существенно предпочтительнее средней арифметической. Показатель **Me** более предпочтителен, чем средне арифметическая, для небольших ранжированных **ВР,** ибо на величину средней могут оказать влияние случайные колебания значений его *крайних* элементов. В случае отличного от *нормального* распределения **ВР** соотношения значений показателей **Mo, Me** и **Ā** можно использовать для характеристики его *асимметрии*.

## 6.4. Метод средних – важный прием обобщения

***Метод средних*** настолько важен в статистике, что его иногда неправомерно ассоциируют с самой статистикой (*например*, *А.Л. Боули*). Однако метод средних наряду с положительными сторонами имеет достаточно серьезные недостатки и его применение требует определенной осторожности. ***Средняя*** только тогда корректно характеризует типы и закономерности явлений, когда она вычисляется для качественно *однородной* совокупности. При этом, имеется в виду не полное совпадение и тождество всех единиц в совокупности, а общие признаки, позволяющие вычислять корректную среднюю величину. Если средняя вычисляется для качественно разнородной совокупности, то в ней могут затушеваться различные социально-экономические типы явлений. Такая средняя из корректной превращается в фиктивную величину. Примером такого рода может служить, так называемая "*средняя*" зарплата в государстве, вычисляемая без учета групповой дифференциации населения по доходам и расходам.

Наибольшая эффективность метода средних достигается при разбиении совокупности на *однородные* группы, внутри которых вычисляются средние, характеризующие каждую

группу. В средней величине взаимно погашаются крайние значения величин и действия случайных факторов, позволяя более четко проявиться типу и качеству явления. В средней величине все влияния на объект *взаимопогашаются* и более четко выражается основная линия его развития, в ней устраняются все случайные колебания отдельных единиц и отражается объективная необходимость. *Средняя* выступает как величина обобщающая, типическая для данного явления.

***Средняя величина*** – это *обобщающий показатель*, характеризующий *типичный уровень* явления. Он выражает величину признака, отнесенную к единице совокупности. Для того, чтобы этот ***средний показатель*** был действительно типизирующим для исследуемой совокупности, он должен рассчитываться с учетом определенных принципов, из которых отметим следующие четыре, а именно *средняя величина* должна: (**1**) определяться для совокупностей, состоящих из *качественно однородных* единиц, (**2**) вычисляться для совокупностей, состоящих из достаточно большого числа единиц, (**3**) рассчитываться для совокупностей, единицы которых находятся в *нормальном*, естественном состоянии и (**4**) вычисляться с учетом социально-экономического содержания исследуемого показателя.

***Общая теория статистики*** предлагает общие принципы и обоснования основных форм применения средних величин. Тогда как отраслевые статистики имеют дело с конкретным разнообразием средних; они занимаются вопросами анализа средних, особенностями практики их применения и вычисления [231]. Для теории статистики *(особенно отраслевых статистик*) главными являются не математическая сторона вычисления средних, а приемы анализа средних по существу, выявление границ и особенностей использования того или иного технического приема, использование средних для изучения и иллюстрации тех или иных закономерностей социально-экономических явлений.

С помощью ***метода средних*** одним числом характеризуется вся исследуемая совокупность единиц, выявляя общие черты и устраняя случайные. Но это далеко не всегда имеет место, ибо средняя величина лишь до известного предела характеризует качество явлений. Необходимо, чтобы взаимопогашению отклонений от средней в процессе ее вычисления соответствовал реальный процесс, характеризующий явление. В любом случае средняя – элемент абстракции, ибо в ней отсутствуют индивидуальные различия и ее значение часто не является элементом исследуемой совокупности или с реальной точки зрения не имеет смысла. Например, в среднем за год **ТТГ** публиковала по **1.77** монографии, книги или отчета (табл. **7**), хотя эти публикации измеряются только целыми числами. Однако, средняя отражает общее в массе явлений и это общее реально существует. Если же средняя величина вычисляется для произвольной совокупности элементов, то в ней не отразится общего для них качества, как реально не существующего для разнородных явлений и процессов.

Конкретный *экономический анализ* помогает определять допустимость той или иной средней. В любом случае при использовании средней следует иметь в ввиду ее недостатки и, по-возможности, оценивать ее ошибки. В практике статистики широко применяется индексный метод, представляющий собой дальнейшее развитие ***метода средних*** и рассматриваемый в девятой главе настоящей книги.

В заключение обсуждения отметим, что ***средняя*** всегда дает обобщающую характеристику лишь по одному признаку, тогда как каждое явление многогранно – оно характеризуется многими признаками. Поэтому, для более глубокого его анализа рекомендуется вычислять не одну, а некоторые системы средних, позволяющих описывать явление с разных сторон. Более того, сама *система средних* должна применяться в комплексе с другими обобщающими показателями – *объемными* и *относительными* величинами. Лишь в этом случае имеется определенная гарантия того, что совокупность явлений познается глубоко и всесторонне.

# Глава 7.
# Элементы анализа вариационных рядов

Как уже говорилось выше, ***вариация признака*** – *изменение его значений* у единиц *совокупности*. Элементы совокупности характеризуются различными *качественными* и *количественными* признаками. Например, элементы совокупности научных публикаций **ТТГ** (табл. **1, 3, 7**) характеризуются такими признаками, как тип и место (*качественные*); время, количество и объем (*количественные*) публикаций. Можно было бы активно использовать для анализа совокупности и такой важнейший количественный признак, как частота цитирования в отечественной и зарубежной литературе, но по целому ряду причин полностью достоверных первичных данных по данному признаку получено не было. Однако в анализе полученные данные используются. Подсчитав число одинаковых значений признака у элементов данной совокупности, получаем ряд распределения – *вариационный ряд* (**ВР**), в котором отдельные значения признака называются ***вариантами***. В западной статистической литературе для *вариационный ряда* часто используется термин "*упорядоченная выборка*", однако, на протяжении настоящей книги мы будем использовать именно первый термин, широко используемый в российской статистической литературе. Естественно, это абсолютно не влияет на общность рассмотрения всего последующего материала.

***Вариационный ряд*** (**ВР**) – упорядоченное в порядке возрастания множество всех значений случайной выборки **X = {x1, x2, ..., xn}** с функцией распределения **F(X)**; при этом, **j**-й член ряда называется **j**-й *порядковой статистикой*, а его номер – *рангом* статистики. **ВР** служит основой для построения эмпирической функции распределения статистических данных, а именно: **Fn(X) = m(x)/n,** где **m(x)** – число членов ряда, меньших **X**-величины. **ВР** находят весьма широкое применение при первичной статистической обработке данных, в частности, при сравнении уровней экономических показателей различных объектов (*отраслей*, *корпораций*, *предприятий и т.д.*).

Количественная сторона явлений выступает как *переменная величина*, имеющая определенные границы. *Вариация* порождается целым комплексом разнообразных условий, действующих на элементы совокупности. Так, например, на вариации признаков элементов-публикаций влияют такие условия, как: творческие способность и активность, изменение тенденций в мировой науке, возможности издания, социально-экономические условия и др. Изучение вариации наряду с применением методов *средних* и *относительных* величин имеет большое практическое и методологическое значения. ***Вариация*** характеризует *степень однородности* совокупности по данному признаку. Таким образом, ***вариация,*** т.е. несовпадение уровней одного и того же показателя у разных объектов, имеет объективный характер и помогает познать сущность изучаемого явления. *Вариация* присуща всем явлениям природы и общества, кроме законодательно закреплённых нормативных значений отдельных социальных признаков.

Измерение вариации позволяет определять степень воздействия на данный признак других варьирующих признаков; оно также необходимо при организации *выборочного* наблюдения, при изучении статистических взаимосвязей и в ряде других случаев. Особое значение приобретают вариации по времени, ибо характеризуют динамику изучаемого объекта.

## 7.1. Показатели вариации совокупностей

Для выявления закономерности **ВР** недостаточно его графического изображения, поэтому для анализа **ВР** и *сравнительной* характеристики рядов применяются *обобщающие* показатели. Одну из групп таких показателей составляют характеристики *центра группирования* (**Ā, Mo** и **Me**), рассмотренные в предыдущей главе. Вторую группу составляют показатели *степени вариации*, характеризующие колеблемость признака; третья группа – показатели *формы* **ВР**. *Первая* группа показателей **ВР** рассмотрена в предыдущей главе (*раздел* **6.1**) настоящей книги.

Показатели ***степени вариации*** относятся к числу *обобщающих* показателей, измеряя вариацию в совокупности явлений. В статистической практике чаще всего используются следующие показатели вариации, а именно: *размах вариации* (**R**), *среднее абсолютное отклонение* ($\overline{\mathbf{D}}$), *среднее квадратическое отклонение* (**σ**), *дисперсия* ($\boldsymbol{\sigma^2}$) и *коэффициент вариации* (**V**). Значение указанных показателей вариации состоит в том, что они:

- *дополняют средние величины, скрывающие индивидуальные различия единиц совокупности*
- *характеризуют степень однородности статсовокупности по данному признаку*
- *характеризуют границы вариации данного признака*
- *соотношения показателей вариации характеризуют взаимосвязь между выбранными для изучения признаками явления.*

Рассмотрим показатели вариации детальнее. **ВР** имеет в общем случае вид **X = {x1, x2, ..., xn}.** *Размах вариации* (**R**) – разность между максимальным и минимальным значениями признака **X,** т.е. **R = (Xmax - Xmin).** Показатель **R** характеризует пределы изменения варьирующего признака; его простота и определяет частое использование в технике и экономике. Однако, величина **R** зависит только от двух крайних значений **ВР,** что в известной мере делает его величину случайной. Более надежным является *средний размах* ($\overline{\mathbf{R}}$), вычисляемый как *средняя* арифметическая из ряда размахов, полученных в результате ряда серий наблюдений. Так, показатель $\overline{\mathbf{R}}$ используется при контроле *качества продукции*, особенно на *поточных* линиях.

***Среднее абсолютное отклонение*** ($\overline{\mathbf{D}}$) вычисляется по следующим формулам:

$$\overline{\mathbf{D}} = \frac{\sum_{j=1}^{n} |x_j - \overline{X}|}{n} \qquad \widetilde{\mathbf{D}} = \frac{\sum_{k=1}^{m} |x_k - \overline{X}| f_k}{\sum_{k=1}^{m} f_k}; \qquad \sum_{k=1}^{m} f_k = n \tag{42}$$

где вторая формула определяет *средневзвешенное* значение $\widetilde{\mathbf{D}}$, $\mathbf{f_k}$ – частота (*вес*) варианты $\mathbf{x_k}$ **X**-переменной (**k = 1 .. m**). Данный показатель используется, например, для характеристики однородности пряжи в текстильной промышленности. ***Среднее квадратическое отклонение*** (**σ, СКО**) вычисляется по следующим общим формулам:

$$\sigma = \sqrt{\frac{\sum_{j=1}^{n} (x_j - \overline{X})^2}{n}} \qquad \widetilde{\sigma} = \sqrt{\frac{\sum_{k=1}^{m} (x_k - \overline{X})^2 f_k}{\sum_{k=1}^{m} f_k}} \tag{43}$$

где вторая формула определяет средневзвешенную величину **СКО** ($\widetilde{\sigma}$); $\mathbf{f_k}$ – частота (*вес*) варианты $\mathbf{x_k}$ **X**-переменной (**k = 1 .. m**). Этот показатель является наиболее фундаментальным при определении вариации признака и весьма широко применяется не только в статистике, но и в других сферах человеческой деятельности. Приведем примеры расчета введенных

показателей вариации на основе **ВР** (41), определенного в разделе **6.3,** сведя промежуточные вычисления в табл. **8.**

**Таблица 8. Разработочная таблица для вычисления показателей вариации**

| *Интервал* | *Частота* $f_k$ | $x_k$ | $\lvert x_k-\overline{X}\rvert$ | $(x_k-\overline{X})^2$ | $\lvert x_k-\overline{X}\rvert * f_k$ | $(x_k-\overline{X})^2 * f_k$ | $x_k * f_k$ |
|---|---|---|---|---|---|---|---|
| **0 - 1** | **6** | **0** | **1.77** | **3.13** | **10.62** | **18.78** | **0** |
| **1 - 2** | **9** | **1** | **0.77** | **0.59** | **6.93** | **5.31** | **9** |
| **2 - 3** | **7** | **2** | **0.23** | **0.05** | **1.61** | **0.35** | **14** |
| **3 - 4** | **4** | **3** | **1.23** | **1.51** | **4.92** | **6.04** | **12** |
| **4 - 5** | **2** | **4** | **2.23** | **4.97** | **4.46** | **9.94** | **8** |
| **5 - 6** | **2** | **5** | **3.23** | **10.43** | **6.46** | **20.86** | **10** |
| ***Итого:*** | **30** | **15** | **9.46** | **20.68** | **35.00** | **61.28** | **53** |

где $\overline{X} = 1.77$. С учетом данных табл. **8** и вышесказанного довольно легко получаем следующие значения показателей вариации ранее определенного **ВР** (табл. **7**), а именно**:** $\tilde{D} =$ 1.17, $\tilde{\sigma} =$ 1.43, **R** = (5 - 0) = 5. Очевидно, чем меньше величина **СКО,** тем однороднее совокупность и выше ее качество. Тогда как для явно несимметричных распределений вычисление **СКО** не имеет особого смысла. Между величинами $\tilde{D}$ и $\tilde{\sigma}$ существует приблизительное соотношение $\tilde{\sigma} \approx 1.25 * \tilde{D}$**,** если фактическое распределение близко к нормальному. Для нашего примера (41) получаем оценку **АК =** $\tilde{\sigma}/\tilde{D}$ = 1.43/1.17 = 1.22**,** т.е. $\lvert$**АК -** 1.25$\rvert$*100**/R** = 0.6%**,** что является вполне удовлетворительной близостью для такого рода совокупности.

В статистике широко используется показатель вариации, называемый ***дисперсией***. В западной статистической литературе для данного показателя часто используется термин *"вариация"***;** однако, на протяжении настоящей книги наряду с этим термином мы будем использовать термин *"дисперсия"*, весьма широко используемый в российской статистической литературе. Естественно, это абсолютно не нарушает общность рассмотрения последующего материала. ***Дисперсия*** вычисляется по следующей достаточно простой формуле:

$$\sigma^2 = \frac{\sum_{j=1}^{n}(x_j-\overline{X})^2}{n} \qquad \tilde{\sigma}^2 = \frac{\sum_{k=1}^{m}(x_k-\overline{X})^2 f_k}{\sum_{k=1}^{m} f_k} \qquad \sigma^2 = \frac{\sum_{j=1}^{n} x_j^2}{n} - \overline{X}^2 \tag{44}$$

Третья формула в (44) представляет *упрощенный* вариант вычисления дисперсии. Вычисление средневзвешенной дисперсии ($\tilde{\sigma}^2$) не составляет особого труда, если используются частоты вариантов признака (*подобно случаю средних величин*). Вторая формула (44) представляет собой *взвешенную дисперсию*. Как правило, вычисление дисперсии предшествует вычислению **СКО,** но дисперсия имеет и вполне самостоятельное значение, о чем будет идти речь ниже.

Для упрощения вычисления дисперсии можно использовать ее основное свойство: *изменение всех значений признака на величину* **p** (*в* **p** *раз*) *не изменяет* (*изменяет в* $\mathbf{p^2}$ *раз*) *величину дисперсии*. Ряд других полезных свойств дисперсии приведен в разделе **2.3** книги. Для случая примера (41) значение дисперсии равно $\tilde{\sigma}^2$ = 1.56. В качестве еще одного примера вычислим *дисперсию* для *альтернативных признаков*. *Альтернативный признак* имеет только два значения, а именно: **1** (*наличие свойства*) и **0** (*отсутствие его*). Пусть в некоторой совокупности **X** *доли* признаков со

значениями "**1**" и "**0**" соответственно равны **P** и **Q** (**P + Q = 1**). Тогда на основе вышесказанного получаем следующие вполне очевидные соотношения:

$$P + Q = 1; \qquad \overline{X} = \frac{1*P + 0*Q}{P+Q} = \frac{P}{P+Q} = p \qquad \overline{X^2} = \frac{1^2*P + 0^2*Q}{P+Q} = p$$

$$\sigma^2 = \overline{X^2} - (\overline{X}) = P - P^2 = P*(1-P) = P*Q$$

Данный пример может быть полезен при исследовании вариации качественных признаков. Например, в результате приемки из 5000 готовых *интегральных схем* (**ИС**) лишь 125 оказались бракованными. Используя предыдущие соотношения, легко получаем:

| *Значение переменной* | *Кол-во таких значений* |
|---|---|
| **1** | **4875** |
| **0** | **125** |
| **Итого:** | **5000** |

Если **P** – доля пригодных **ИС** и **Q** – доля бракованных **ИС**, то простой подсчет нам дает:

$$\overline{X} = P = (1*4875 + 0*125)/5000 = 0.975, \quad Q = 1 - P = 1\text{-}0.975 = 0.125$$

*Дисперсия* и **СКО** доли бракованных **ИС** равны **P*Q** = 0.122 и $\sqrt{P*Q}$ = 0.349 соответственно.

***Дисперсия*** и **СКО** – наиболее широко используемые показатели *степени вариации* признаков, ибо они входят в большинство теорем теории вероятностей, служащих фундаментом математической статистики. Наряду с этим, дисперсия разлагается на составные элементы, позволяющие оценивать влияние различных факторов на вариацию признака.

При вычислениях суммирующих показателей для интервального **ВР** значения переменной замененяются центральными значениями интервала, которые вообще говоря, являются отличными от средних интервальных. В результате, при вычислении дисперсии возникает *систематическая* ошибка. В. Шеппард нашел, что ошибка дисперсии, обусловленная этим обстоятельством составляет **1/12** квадрата *интервального значения* (**ИЗ**), то есть исправленная *дисперсия* вычисляется как $\sigma^2 - IV^2/12$. Поправку Шеппарда следует применять при условиях, что: (**1**) **ВР** имеет непрерывный характер вариации, (**2**) характеризуется близостью к **X**-оси на концах кривой распределения, и (**3**) имеет достаточно большое количество данных (**n ≥ 500**).

Рассмотренные выше показатели вариации, исключая дисперсию, выражаются в единицах измерения признака и являются абсолютными. Поэтому **СКО** неудобно для сопоставления вариации различных признаков. Для этих целей используются *коэффициенты вариации* (**КВ**), определяющие относительные меры вариации признаков и вычисляемые по следующей принципиальной формуле:

$$VC = \frac{\textit{Абсолютный показатель вариации}}{\textit{Средняя или ее заменяющая}}$$

Примером таких формул могут служить следующие показатели вариации:

$$V = \frac{\sigma*100\%}{\overline{X}} \qquad V1 = \frac{\overline{D}*100\%}{\overline{X}} \qquad V2 = \frac{\sigma*100\%}{M_0} \qquad V3 = \frac{R*100\%}{\overline{X}} \tag{45}$$

из которых наиболее употребительным является **КВ V.** При этом, вместо величин $\sigma$ и $\overline{D}$ для случая интервальных **ВР** используются их *взвешенные* аналоги $\tilde{\sigma}$ и $\tilde{D}$ (42, 43). Вычислим все четыре коэффициента (45) для нашего примера интервального ряда (41):

**V** = (1.43/1.77)*100 = 80.8% **V1** = (1.17/1.77)*100 = 66.1%
**V2** = (1.43/1.6)*100 = 89.48% **V3** = (5/1.77)*100 = 282.5%

При этом, значения показателей **σ, $\overline{X}$, $\overline{D}$, R** и **Mo** были получены нами выше.

Посредством показателя **KB** (45) можно сравнивать *размеры* вариации идентичных признаков в различных совокупностях либо различных в одной и той же совокупности. Показатель **KB V** используется не только для сравнительной оценки, но и как характеристика однородности совокупности. Для распределений, довольно близких к *нормальному*, совокупность полагается однородной, если величина **V ≤** 33%. В случае нашего примера (41) **V =** 80.8% **>>** 33%, т.е., на *первый взгляд*, исследуемая совокупность не должна полагаться *однородной*. С другой стороны, в разделе **6.3** показано, что для нее получаем $\overline{X}$ = 1.77, **Mo** = 1.6 и **Me** =2, т.е. в предположении нормального распределения, совокупность может полагаться однородной. Полученное нами противоречие вполне можно интерпретировать как отличие распределения совокупности от *нормального*, что *согласуется* со здравым смыслом и самой сутью *явления*. Ниже мы продолжим исследование формы данного распределения более детально.

## 7.2. Меры вариации сгруппированных данных совокупности

Для оценки влияния различных факторов, определяющих вариацию индивидуальных значений признака, можно воспользоваться разложением дисперсии на составляющие: *межгрупповую* ($\delta^2$) и *внутригрупповую* ($\sigma_j^2$) дисперсии. *Общая* дисперсия $\sigma^2$ измеряет *вариацию* признака во всей совокупности под влиянием всех факторов. ***Внутригрупповая*** дисперсия ($\sigma_j^2$) измеряет вариацию признака внутри самой **j**-группы совокупности м вычисляется как:

$$\sigma_j^2 = \sum_{k=1}^{n_j} \frac{(x_{kj} - \overline{X}_j)^2}{n_j} \qquad \sum_{j=1}^{m} n_j = n \tag{46}$$

где $\overline{X}_j$ – средняя по **j**-группе совокупности. Межгрупповая дисперсия $\delta^2$ измеряет вариацию *групповых средних* $\overline{X}_j$ относительно *общей средней* $\overline{X}$ совокупности и вычисляется как:

$$\delta^2 = \frac{\sum_{j=1}^{m} (\overline{X}_j - \overline{X})^2}{m} \tag{47}$$

*Общая дисперсия* измеряет степень вариации признака, порождаемой всей совокупностью действующих на него факторов. *Межгрупповая дисперсия* измеряет *степень вариации* признака в совокупности за счет фактора, положенного в основу группировки – группировочного признака. Тогда как *внутригрупповая дисперсия* измеряет влияние на вариацию всех прочих факторов, кроме *группировочного*. Три вида дисперсии подчиняются общему *правилу сложения дисперсий*, а именно:

$$\sigma^2 = \delta^2 + \overline{\sigma}_j^2 \tag{48}$$

т.е. *общая дисперсия равна сумме среднегрупповой и межгрупповой дисперсий*. Его суть состоит в том, что общая дисперсия, возникающая под влиянием всех факторов, состоит из суммы дисперсий, возникающих под влиянием группировочного фактора и всех прочих факторов. По любым двум видам дисперсии, таким образом, можно найти третий. На основе данного правила можно удобно оценивать удельное влияние группировочного фактора во всей совокупности факторов, действующих на результативный признак.

В статистике также рассматривается еще один показатель – ***корреляционное отношение*** (**KO**), вычисляемое по следующей простой формуле:

$$\eta = \sqrt{\frac{\delta^2}{\sigma^2}} = \sqrt{\frac{\delta^2}{\sigma^2 + \overline{\sigma}_j^2}} = \sqrt{\frac{1}{1 + \overline{\sigma}_j^2 / \delta^2}} \qquad (49)$$

**КО** (**η**) характеризует *степень* влияния *группировочного* признака на *результативный*. *Значение* **КО** удовлетворяет условию $\mathbf{1 \geq \eta \geq 0}$ и определяет влияние *группировочного* признака (*фактора*) от максимального при $\mathbf{\eta = 1}$ до его полного отсутствия при $\mathbf{\eta = 0}$; для его вычисления нужно знать любые два вида дисперсии. Для иллюстрации вышесказанного рассмотрим пример, базирующийся на табл. **9,** впоследствии используемой нами при изучении элементов *анализа* временных (*динамических*) рядов.

**Таблица 9. Распределение ежегодных ссылок на публикации ТТГ (1970 – 1999)**

| *Год* | *Количество публикаций* | | *Количество ссылок* | | *Всего* | |
|---|---|---|---|---|---|---|
| | *отечественные* | *зарубежные* | *отечественные* | *зарубежные* | *работ* | *ссылок* |
| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1970 | 3 | 0 | 0 | 0 | 3 | 0 |
| 1971 | 9 | 0 | 0 | 33 | 9 | 33 |
| 1972 | 4 | 0 | 0 | 38 | 4 | 38 |
| 1973 | 2 | 0 | 5 | 42 | 2 | 47 |
| 1974 | 3 | 3 | 7 | 76 | 6 | 83 |
| 1975 | 2 | 1 | 12 | 94 | 3 | 106 |
| 1976 | 4 | 1 | 13 | 96 | 5 | 109 |
| 1977 | 0 | 0 | 14 | 98 | 0 | 112 |
| 1978 | 1 | 1 | 14 | 102 | 2 | 116 |
| 1979 | 0 | 1 | 15 | 125 | 1 | 140 |
| 1980 | 3 | 0 | 18 | 145 | 3 | 163 |
| 1981 | 6 | 2 | 27 | 174 | 8 | 201 |
| 1982 | 3 | 2 | 24 | 186 | 5 | 210 |
| 1983 | 4 | 1 | 22 | 205 | 5 | 227 |
| 1984 | 0 | 3 | 29 | 220 | 3 | 249 |
| 1985 | 4 | 1 | 42 | 222 | 5 | 264 |
| 1986 | 1 | 3 | 48 | 236 | 4 | 284 |
| 1987 | 1 | 2 | 56 | 240 | 3 | 296 |
| 1988 | 3 | 8 | 64 | 266 | 11 | 330 |
| 1989 | 1 | 2 | 68 | 284 | 3 | 352 |
| 1990 | 2 | 1 | 68 | 292 | 3 | 360 |
| 1991 | 3 | 1 | 70 | 304 | 4 | 374 |
| 1992 | 1 | 0 | 65 | 314 | 1 | 379 |
| 1993 | 3 | 0 | 58 | 316 | 3 | 374 |
| 1994 | 1 | 0 | 55 | 320 | 1 | 375 |
| 1995 | 1 | 1 | 60 | 336 | 2 | 396 |
| 1996 | 2 | 1 | 63 | 345 | 3 | 408 |
| 1997 | 3 | 2 | 67 | 384 | 5 | 455 |
| 1998 | 5 | 5 | 75 | 395 | 10 | 470 |
| 1999 | 5 | 5 | 95 | 405 | 10 | 500 |
| Итого | 80 | 47 | 1154 | 6293 | 127 | 7451 |

При сводке результатов наблюдения относительно цитируемости работ **ТТГ** учитывались только доступные в библиотеках АН Эстонии (Таллинн), ТГУ (Тарту), ГПНТБ, б-ка им. В.И. Ленина, БЕН АН СССР (Москва), АН Украины (Киев), Литвы, Латвии и др. периодические, непериодические и реферативные издания (*отечественные* и *зарубежные*), а также *информация* из некоторых баз данных ряда ведущих научных и информационных центров США, Японии и Германии. При этом, ссылки авторов на собственные работы не учитывались. Несмотря на вполне естественную неполноту данных данного статистического наблюдения, результаты сводки (табл. **9**) позволяют проводить интересный статистический анализ цитируемости научных работ **ТТГ** в области теории однородных структур (*Cellular Automata*). На основе данных табл. **9** делаем сводку публикаций **ТТГ** по пятилеткам ее деятельности, включая их группировку по признаку – место издания (*за рубежом*, *СССР*). Результатом сводки является следующая сводная табл. **10.**

**Таблица 10. Распределение публикаций ТТГ по однородным структурам по пятилеткам**

| *Период* / *Место* | 1970 - 1979 | 1975 - 1979 | 1980 - 1984 | 1985 - 1989 | 1990 - 1994 | 1995 - 1999 | *Всего* | *Совокупность Группы* **А** *и* **В** |
|---|---|---|---|---|---|---|---|---|
| ***Отечественные*** | **21** | **7** | **16** | **10** | **10** | **16** | **80** | **A** |
| ***Зарубежные*** | **3** | **4** | **8** | **16** | **2** | **14** | **47** | **B** |
| ***Итого:*** | **24** | **11** | **24** | **26** | **12** | **30** | **127** | **C** |

Перед дальнейшим рассмотрением сделаем одно замечание терминологического характера. Статистические совокупности и ряды, изучаемые в иллюстративных целях, базируются на данных, составляющих базу данных публикаций **ТТГ** в области *Математической Теории Однородных Структур* (**МТОС**) и ее приложений. База данных формировалась в течение 1970 – 2000 годов. В то время как, в английской научной литературе для данного направления используется термин "*Cellular Automata*" (*Клеточные Автоматы*), полностью эквивалентный термину "*Однородные Структуры*"**.** Поэтому, не нарушая общности, мы будем предполагать ниже, что оба эти термина полностью идентичны.

Целью нашего анализа будет совокупность **С,** определяющая число публикаций **ТТГ** в области **МТОС** по пятилеткам в целом, и составляющие ее подгруппы **А** и **В** публикаций, изданных соответственно в СССР и за рубежом. Для этого вычисляем по совокупности и ее подгруппам *средние*, *дисперсии*, *межгрупповую дисперсию* и *корреляционное отношение*, а именно:

$$\mathbf{C = \frac{127}{12} = 10.53} \qquad \mathbf{A = \frac{80}{6} = 13.33} \qquad \mathbf{B = \frac{47}{6} = 7.83} \qquad \mathbf{\sigma_A^2 = 22.56}$$

$$\mathbf{\sigma_B^2 = 29.51} \qquad \mathbf{\sigma_C^2 = 33.58} \qquad \mathbf{\sigma_C^2 = \delta_{AB}^2 + \frac{\sigma_A^2 + \sigma_B^2}{2}} \qquad \mathbf{\eta = \sqrt{\frac{7*54}{33.58}} = 0.474}$$

Читателю рекомендуется в качестве весьма полезного упражнения проверить полученные результаты. По приведенным данным получаем, что фактор места издания, положенный в основу группировки (табл. **10**), достаточно существенно влияет на число публикаций, но существует и ряд других факторов, влияние которых также достаточно существенно на результативный признак – количество публикаций **ТТГ** в данной области кибернетики.

Изучение вариации имеет особый смысл только в пределах однородной совокупности. Пределы вариации дают представление о тех границах применения значений признака, за которыми следуют качественные изменения. Важное значение имеет также закономерность

случайной вариации. По нормальному закону распределения вариация значений признака находится в пределах $\{\overline{X} - 3*\sigma, \overline{X} + 3*\sigma\}$ – так называемое *правило трех сигм* (*раздел* **2.5**). Между тем, действие этого закона ограничено, ибо для него необходимо постоянство *среднего уровня* и отсутствие доминирующих факторов, влияющих на вариацию средней.

Однако, в ряде случаев использование правила трех сигм практически возможно. Например, для случая примера 41) имеем: $\overline{X}_1 = 1.77$ и $\sigma_1 = 1.43$. Следовательно, число публикаций **ТТГ** должно варьироваться в пределах от $(\overline{X}_1 - 3*\sigma_1)$ до $(\overline{X}_1 + 3*\sigma_1)$, т.е. в интервале (**0-6**), что вполне отвечает реалиям. Второй пример (табл. **9,** графа **1**) дает следующие результаты: $\overline{X}_2 = 2.67$, $\sigma_2 = 1.94$; вариация числа публикаций находится в интервале $[0 - (\overline{X}_2 + 3*\sigma_2)] = (0 - 9)$, т.е. и в этом случае *правило трех сигм* дает правильные результаты, хотя распределение значений признака и не является *нормальным*. В качестве достаточно полезного упражнения читателю рекомендуется проверить справедливость данного правила и для совокупностей данных из граф **3, 6** (см. табл. **9**).

Вариация *средних* меньше вариации значений признака и изменяется в пределах $\{\overline{X} - 3*\sigma/\sqrt{n}, \overline{X} + 3*\sigma/\sqrt{n}\}$, где **n** – число единиц совокупности. Для приведенных выше двух примеров мы получаем следующие пределы вариации средних величин, а именно:

$$\{\overline{X}_1 - 3*1.43/\sqrt{6}, \overline{X}_1 + 3*1.43/\sqrt{6}\} = \{\overline{X}_1 - 1.751, \overline{X}_1 + 1.751\}$$

$$\{\overline{X}_2 - 3*1.94/\sqrt{30}, \overline{X}_2 + 3*1.94/\sqrt{30}\} = \{\overline{X}_2 - 1.063, \overline{X}_2 + 1.063\}$$

т.е. диапазоны вариации средних значительно уже диапазонов вариации индивидуальных значений признаков исследуемой совокупности.

## 7.3. Элементы анализа формы кривой распределения

Законы распределения являются обобщающей характеристикой вариации в однородной совокупности. *Эмпирическое* распределение может быть графически представлено некоторой кривой, когда ось абсцисс отводится для *вариантов* признака, а ось ординат – их *частотам* (*частостям*). Для получения приблизительного представления о форме такого *распределения* строят его ***график*** (*полигон* и/или *гистограмму*). Так как число *эмпирических* данных, обычно, невелико, то распределение, построенное на их основе, подвержено влиянию случайностей. С увеличением числа данных и уменьшением величины интервала это влияние *уменьшается*, а *эмпирический* полигон сглаживается и переходит в гладкую кривую распределения. Данная кривая характеризует ***теоретическое распределение,*** получающееся при полном погашении всех случайных факторов, затеняющих основную закономерность исследуемого явления.

Статистика имеет дело с различными распределениями; при однородных совокупностях, как правило, имеем дело с *одновершинными* (*унимодальными*; **U-***распределениями*) распределениями. Многовершинность говорит о неоднородности совокупности и в этом случае, как правило, требуется ее перегруппировка с целью выявления *более однородных* групп. Выяснение общего характера распределения **X** предполагает оценку степени его однородности и основных характеризующих его показателей: $\overline{X}$, **Mo, Me, R,** $\overline{D}$, $\sigma$, $\sigma^2$, **V,** коэффициентов *асимметрии* и *эксцесса*. Оценка *эмпирического* распределения сводится к решению трех основных задач:

1. *нахождение общего характера эмпирического распределения* **E(X)**
2. *выравнивание (сглаживание) распределения* **E(X)** *посредством теоретического распределения* **T(X),** *получаемого по тому или иному способу*
3. *оценка степени соответствия распределений* **E(X)** *и* **T(X)** *с помощью критериев согласия гипотез, т.е. проверка степени их близости*.

Практика подсказывает наиболее эффективные способы решения данной задачи. На *первом* этапе на основе первичных или вторичных данных формируется совокупность значений признака, распределение которых предстоит исследовать. Затем строятся гистограмма и полигон частот (*частостей*) на выбранных интервалах [*эмпирическое* **E(X)**-*распределение*] на предмет выбора *наиболее* приемлемого (*по визуальной оценке*) *теоретического* **T(X)**-*распределения*. Форма полигона может служить хорошей отправной точкой для данной цели при условии соблюдения масштаба в выбранной системе координат, как правило, декартовой.

В первую очередь следует обратить внимание на уже известные *теоретические распределения* (раздел **2.4;** *источники* [55,58,62,63,94,96,115,125,132,155-157,161,178,182,270-272,279-281] *и др.*), и, в первую очередь, на нормальное распределение, имеющее место для многих однородных совокупностей. После этого вычисляются упомянутые показатели для **E(X)**-распределения, позволяющие уточнить требования к искомому *теоретическому* **T(X)**-распределению.

На *втором* этапе, базируясь на результатах предыдущего, для *эмпирического* **E(X)**-*распределения* определяем тем либо иным способом *теоретическое* **T(X)**-*распределение*, выравнивающее *первое*. В случае же невозможности отыскать *подходящий аналог* **T(X)**-распределения среди уже *известных* (*либо непосредственно*, *либо путем преобразования*: *сжатие/растяжение*, *сдвиги по осям координат и др.*) оно определяется другим путем: *методом наименьших квадратов*, *графическим способом и др*. Во многом, способ нахождения **T(X)**-распределения зависит от характера данных и навыков самого исследователя.

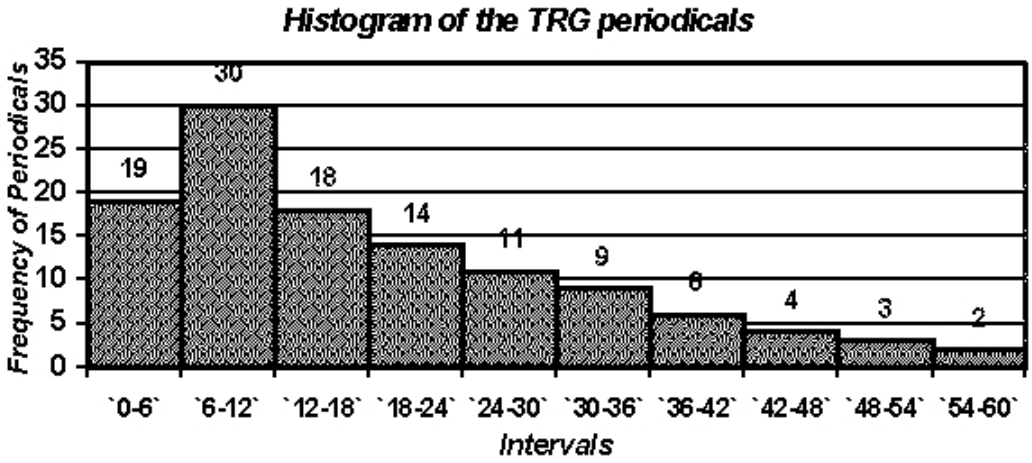
Наконец, *третий* этап состоит в верификации соответствия найденного **T(X)**-распределения эмпирическому **E(X)**. Степень расхождения (*соответствия*) теоретических и эмпирических частот оценивается на основе критериев согласия. Как правило, для этих целей используется критерий согласия К. Пирсона (*критерий* $\chi^2$-*квадрат*). Могут использоваться и ряд других критериев согласия, например, А.Н. Колмогорова, Колмогорова-Смирнова, В. Романовского, Б. Ястремского, Стьюдента (**t**-*критерий*) и др. [96,132,157,159,160,183]. Методику выравнивания **E(X)**-распределений рассмотрим на следующем достаточно интересном примере. Методика базируется на использовании графического метода для получения приближенных с той или иной степенью точности функциональных зависимостей (*т.н. метод компьютерной подгонки*), описывающих статистические явления, которые впоследствии могут быть успешно использованы или в качестве основы для дальнейшей разработки функциональных связей, или в качестве приемлемого окончательного решения.

В качестве первичных данных выбирается множество всех публикаций **ТТГ** за период 1970 – 1999 годы ([23, 26, 28, 29, 190, 191], табл. **7**) и на его основе формируется совокупность **X** (*по признаку – объем публикации в стр.*) всех *периодических* публикаций по **МТОС** и ее *приложениям*. Результаты сводки данной совокупности определяются соответствующим *интервальным* **ВР** (табл. **11**).

**Таблица 11. Распределение периодических публикаций ТТГ (*по объему в страницах*)**

| ***Интервал*** | **0-6** | **6-12** | **12-18** | **18-24** | **24-30** | **30-36** | **36-42** | **42-48** | **48-54** | **54-60** | ***Итого*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Средняя*** | **3** | **9** | **15** | **21** | **27** | **33** | **39** | **45** | **51** | **57** | **300** |
| ***Частота*** | **19** | **30** | **18** | **14** | **11** | **9** | **6** | **4** | **3** | **2** | **116** |

На основе данных табл. **11** строим гистограмму (*histogram*) и полигон (*polygon*) эмпирического распределения (рис. **10, а**). По форме полигона **E(X)**-распределение несколько напоминает нормальное, но для уточнения вычисляем *основные* показатели соответствующего **ВР X** (табл. **11**). В случае *интервальных* **ВР** значения признака вычисляются как *среднеинтервальные*, а для вычислений составляем расчетную табл. **12,** на основе которой легко вычисляем основные показатели исследуемого **ВР** – периодических публикаций **ТТГ** (*the **TRG** periodicals*).



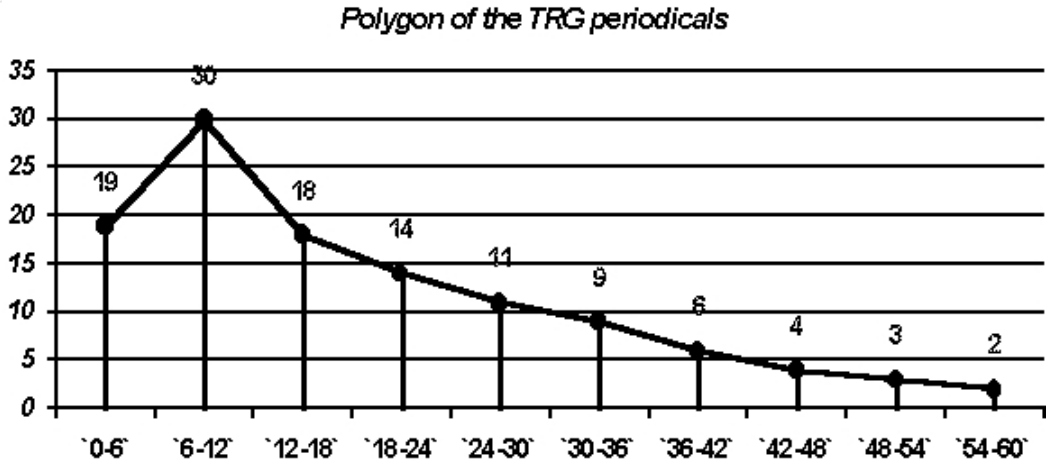




**Рис. 10.а. Эмпирическое E(X)-распределение (*histogram и polygon*), определенное вариационным рядом из табл. 11**

**Таблица 12. Разработочная таблица для интервального ряда X из табл. 11**

| *Интервал* | $x_j$ | $f_j$ | $x_j*f_j$ | $x_j - \overline{X}$ | $\lvert x_j - \overline{X}\rvert f_j$ | $\lvert x_j - \overline{X}\rvert^2 f_j$ | $\lvert x_j - \overline{X}\rvert^3 f_j$ | $\lvert x_j - \overline{X}\rvert^4 f_j$ |
|---|---|---|---|---|---|---|---|---|
| **0 - 6** | **3** | **19** | **57** | **-15.7** | **298.3** | **4683.3** | **-73528.0** | **1154389.1** |
| **6 - 12** | **9** | **30** | **270** | **-9.7** | **291.0** | **2822.7** | **-27380.2** | **265587.8** |
| **12 - 18** | **15** | **18** | **270** | **-3.7** | **66.6** | **246.4** | **-911.8** | **3373.5** |
| **18 - 24** | **21** | **14** | **294** | **2.3** | **32.2** | **74.1** | **170.3** | **391.8** |
| **24 - 30** | **27** | **11** | **297** | **8.3** | **91.3** | **757.8** | **6289.7** | **52204.2** |
| **30 - 36** | **33** | **9** | **297** | **14.3** | **128.7** | **1840.4** | **26317.9** | **376345.4** |
| **36 - 42** | **39** | **6** | **234** | **20.3** | **121.8** | **2472.5** | **50192.6** | **1018909.0** |
| **42 - 48** | **45** | **4** | **180** | **26.3** | **105.2** | **2766.8** | **72765.8** | **1913740.2** |
| **48 - 54** | **51** | **3** | **153** | **32.3** | **96.9** | **3129.9** | **101094.8** | **3265362.1** |
| **54 - 60** | **57** | **2** | **114** | **38.3** | **76.6** | **2933.8** | **112363.8** | **4303532.5** |
| ***Итого:*** | **300** | **116** | **2166** | **113** | **1308.6** | **21727.7** | **267374.9** | **12353835.6** |

Величины **Mo** и **Me** для **ВР X** вычисляются по формулам (39, 40); при этом, интервал (**6 - 12**) ряда **X** является *модально-медианным – мода* и *медиана* принадлежат прилегающим интервалам (**6 - 12**) и (**12 - 18**) соответственно, а именно: **Mo** = 6 + 6*(30 - 19)/(2*30 - 19 - 18) = 8.87, **Me** = 12 + 6*(116/2 - 49)/18 = 15.0. Для других показателей **ВР X,** используя данные табл. **12** и формулы

(42, 43), получаем следующие значения: $\overline{\mathbf{X}}$ = 2166/116 = 18.7, **R** = (57 - 3) =54, $\overline{\mathbf{D}}$ = 1308.6/116 = 11.3, $\sigma^2$ = 21725.7/116 = 187.3, $\sigma$ = 13.7. *Коэффициент вариации* (45) равен **V** = (13.7/18.7)*100 = 73.3%, а сами значения признака лежат в интервале $\{\overline{\mathbf{X}} - 3*\sigma, \overline{\mathbf{X}} + 3*\sigma\}$ = (**0 - 60**).

Выше в разделе **2.3** относительно теоретических распределений рассматривались моменты (*начальные* – $\mathbf{M_j}$ и *центральные* – $\mathring{\mathbf{M}}_\mathbf{j}$), связывающие их *соотношения* (16-18), а также *специальные* коэффициенты *асимметрии* (**A**) и *эксцесса* (**E**). Показатель **A** характеризует "*скошенность*" распределения: при **A > 0** распределение вытянуто от центра ($\overline{\mathbf{X}}$, **Mo**) вправо, при **A< 0** – влево и при **A = 0** – симметрично относительно своего центра.

Используя формулы (16-18) и табл. **12,** для случая нашего примера получаем следующие значения, а именно: $\mathring{\mathbf{M}}_\mathbf{3}$ = 267371.9/116 = 2304.9, $\mathbf{A} = \mathring{\mathbf{M}}_\mathbf{3}/\sigma^3$ = 2304.9/2571.4 = 0.896 **>** 0, т.е. наше экспериментальное **E(X)**-распределение имеет явную *правостороннюю асимметрию*. Показатель **E** характеризует островершинность распределения относительно нормального (*для которого* **E=3**) и вычисляется по следующим простым формулам: : $\mathring{\mathbf{M}}_\mathbf{4}$ = 12353831.6/116 = 106498.5, $\mathbf{E} = \mathring{\mathbf{M}}_\mathbf{4}/\sigma^4 - \mathbf{3}$=106498.5/35227.5 -3=0.023, т.е. распределение **E(X)** более *остро вершинно*, чем *нормальное*, т.к. **E > 0,** и характеризуется скоплением членов **ВР X** в центре распределения. На основе полученных показателей для **E(X)**-распределения можно с уверенностью говорить о том, что оно отлично от нормального. Таким образом, наиболее распространенное (*и, на первый взгляд, наиболее типичное для такого типа явлений*) нормальное распределение следует исключить из дальнейшего рассмотрения.

На следующем этапе на предмет соответствия анализируем другие известные распределения (*например, по справочникам* [55, 93, 94, 162, 186]) и обнаруживаем хорошее визуальное сходство **E(X)**-распределения (рис. **10, а**) и кривой *Гамма*-распределения, имеющего в наиболее общем виде следующую функцию плотности вероятностей:

$$\mathbf{G(x) = \frac{K*b^a*(x-d)^{(a-1)}*e^{-b*(x-d)}}{\Gamma(a)}, \qquad (a, b, K, d, x > 0)} \tag{50}$$

Данное распределение при значении **K = 1** является непрерывным аналогом отрицательного биномиального распределения. Для выравнивания **G(X)**-функцией **E(X)**-распределения необходимо соответствующим образом определить значения ее параметров **a, b, d** и **K.** Для этих целей можно успешно использовать известные классические методы анализа, которые для сложного вида сглаживающих функций достаточно сложны и громоздки. Между тем, с появлением класса **ПК** и развитых программных средств появилась отличная возможность принципиально новой и более эффективной методики выравнивания **ВР,** да и вообще многих эмпирических закономерностей. Для конкретности здесь используется известный математический пакет *Maple*, позволяющий в общепринятой математической нотации решать различные проблемы математического характера в числовом и алгебраическом виде [97, 98, 139, 140, 141, 143, 144, 302, 303]. Детальнее использование данного пакета в контексте его статистических приложений будет рассматриваться в последней главе книги. Здесь же мы приведем документ, написанный в среде *Maple* и решающий нашу задачу выравнивания.

```
> E:= array(1 .. 2, 1 .. 10, [[3, 9, 15, 21, 27, 33, 39, 45, 51, 57], [19, 30, 18, 14, 11, 9, 6, 4, 3, 2]]);
> Fitting:= proc(a, b, K, d, E::array)
  local G, S, Kr, L, H, k, j, p, T, F, z;
  map(with, [plots, linalg]), assign(L = [], F = [TIMES, BOLD, 10]);  z := convert(E, 'listlist');
  T:= textplot({[13, 18, `E(X) ->`], [21.5, 18, `<- T(X)`]});
  G := (x) -> evalf(K*b^a*(x - d)^(a - 1)*exp(1)^(-b*(x - d))/GAMMA(a));
```

```
  H:= add((E[2, j + 1] - G(3 + 6*j))^2/G(3 + 6*j), j = 0 .. 9);
  for k to nops(z[1]) do L:= [op(L), [z[1][k], z[2][k]]] end do;
  S:= listplot(L, 'color' = 'green', 'thickness' = 2);
  Kr:= plot(G(x), x = 1 .. E[1, k - 1], 'thickness' = 2, 'color' = 'magenta'); print(cat(`H=`, convert(H,
  'string'))); display({Kr, T, S}, 'axesfont' = F, 'labels' = [Intervals, Fj], 'labelfont' = F, 'font' = F)
end proc
> Fitting(1.11, 0.055, 750, 2.985, E);
```

**H = 1.484313119**

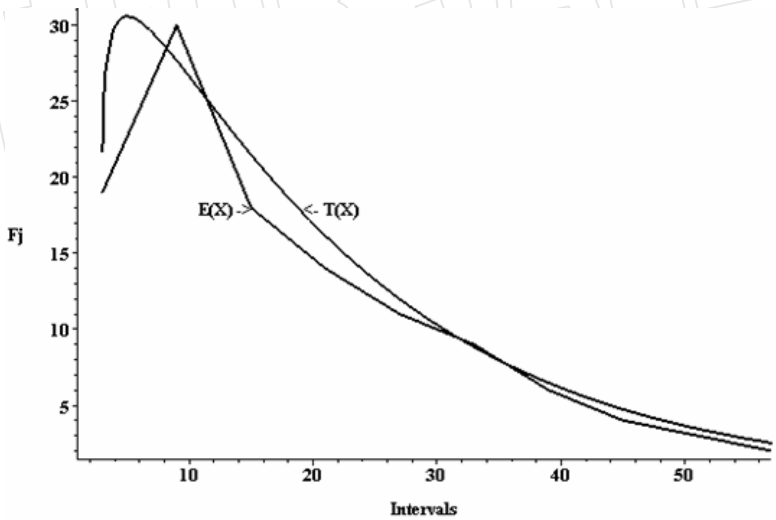



**Рис. 10.b. Сглаживание эмпирического распределения в среде пакета *Maple*, где E(X) – эмпирическое и T(X) – теоретическое распределение, соответствующее E(X).**

В *Maple*-документе (рис. **10.b**) для имеющегося *эмпирического* распределения **E(X)** определена *теоретическая* функция плотности **T(X)** в виде вышеупомянутой функции (50) с получением графиков **E(x)** и **T(x)** в декартовых координатах совместно с возвратом **H**-значения важного показателя $\chi^2$-теста, вычисленного согласно следующей формулы:

$$H = \sum_{k=0}^{9} \frac{(E(x_k) - T(x_k))^2}{T(x_k)}, \quad x_k = 3 + 6*k \ \ (k = 0, 1, 2, \ldots, 9) \tag{51}$$

При этом, графическое представление обоих распределений и вычисление **H**-показателя запрограммированы в *Maple*-процедуре ***Fitting*** с соответствующими аргументами. Путем изменения фактических значений для вышеупомянутых аргументов мы имеем возможность динамически контролировать их влияние на взаимное расположение *эмпирической* кривой **E(X)** и *теоретической* кривой **T(X),** а также значение **H**-*показателя*. Тогда, на основе знания свойств функции **G(X)** (50) и доступных значений $E(x_k)$ мы имеем возможность подбирать значения для параметров **a, b, d** и **K** таким образом, чтобы минимизировать **H**-величину, одновременно визуально наблюдая динамику изменения взаимного расположения кривых **E(x)** и **T(x).** На основе значений **K** = 750, **a** = 1.11, **b** = 0.055 и **d** = 2.985 мы получаем искомое значение показателя: **H = 1.484.** Согласно процедуре Пирсона метод *оценки* степени согласия *теоретического* **T(X)** и *экспериментального* **E(X)** распределений состоит в следующем.

По формуле (51) вычисляется **H**-показатель ($\chi^2$-*тест критерия согласия*), где **E(xk)** и **T(xk)** – частоты экспериментального и теоретического распределения соответственно. Затем на его основе может быть легко рассчитана вероятность **P** согласия обоих распределений. Для вычисления **P**-значения используются специальные таблицы вероятностей **P(H, k)** (*например*,

в [92, 122-125]). Значение вероятности **P** определяется из таблицы **P(H, k)** при условии, что входы в **P**-таблицу – количество степеней свободы **k = n - 4** и **H**-значение, где **n** (*количество интервалов вариационного ряда*) и количество параметров, определяющих теоретическое распределение (50), равняется четырем. На основе **P**-значения судят о важности расхождений между *экспериментальным* **E(X)** и *теоретическим* **T(X)** распределениями. Совпадение считают *хорошим* в случае **P> 0.5,** *приличным* в случае **0.2 ≤ P <0.5,** иначе *неудовлетворительным*.

Для нашего конкретного примера мы получаем следующие значения: **k** = 10 - 4 = 6 и **H** = 1.484, которые позволяют на основе Таблицы VII **P(H, k)** [92] легко вычислять оценки для **P**-величины, а именно: **0.95 <P <0.975.** Для корректировки **P**-значения может использоваться достаточно простая интерполяционная формула:

$$\mathbf{P = (m - d + 1)*P_m^k - (m - d)*P_{m+1}^k} \qquad (52)$$

где (**m, m+1**) – интервал, содержащий значение **H = d,** и $\mathbf{P_m^k}$, $\mathbf{P_{m+1}^k}$ – вероятности для значений **H = m** и **H = m + 1** соответственно при условии, что количество степеней свободы равняется **k.** Наконец, из формулы (52) мы получаем скорректированное значение для **P**-вероятности (*из-за выбранного метода интерполяции даже несколько уменьшенное значение*), а именно: **P** = (1.24 + 1 - 1.484)*0.975 - (1.24 - 1.484)*0.95 = 0.969.

Таким образом, между распределениями **E(X)** и **T(X)=G(X)** существует достаточно хорошее согласие. Это говорит о том, что основу распределения периодических публикаций **TTГ** по их объему составляет именно *Гамма*-распределение. Данный факт интересен и сам по себе, предполагая дальнейшие исследования в этом направлении. Сразу же его можно объяснить тем, что выбранная совокупность содержит как сугубо математические работы, занимающие, как правило, небольшой объем, так и работы прикладного и обзорного характера, объемы которых отличаются большими значениями и их вариацией. В свете полученного результата можно ставить вопрос о своего рода *неоднородности* публикаций относительно их характера: *теоретические* и *прикладные*. В любом случае – это весьма интересный вопрос для *дальнейшего* исследования по статистике научной активности, представляющей несомненный интерес для современного науковедения.

Используя описанную компьютерную методику, нами было проведено некоторое *улучшение* сглаживания **E(X)**-распределения посредством следующей функции:

$$\mathbf{T1}(X) = \mathbf{G}(X) + \mathbf{S}(X),\ \mathbf{S}(X) = 5.375000000\ c\ e^{(-c|x-9|)}$$

где **c = 0.93,** а **S(X)**-функция есть не что иное, как функция плотности распределения Лапласа [55]. Функция **T1(X)** дает существенно лучшие параметры согласия с **E(X)**-распределением, а именно: **H1=1.798** и **P1=0.933 >> 0.5**. Читателю рекомендуется проверить этот факт в качестве полезного упражнения в среде пакета *Maple* или другого подобного ему средства.

Используя вышеупомянутую компьютерную процедуру, мы провели сглаживание целого ряда распределений. Читателю рекомендуют исследовать данную процедуру в среде пакета *Maple* или другого подобного средства в качестве весьма полезного упражнение. В частности, в книгах [14, 15, 29, 127, 190] мы рассмотрели подобную задачу в среде пакета *MatCAD*.

На основе полученного **T1(X)**-распределения можно сделать ряд интересных выводов по **TTГ(X)**-распределению *периодических* публикаций **TTГ** согласно их объема. Подобно большинству социально-экономических явлений **TTГ(X)**-распределение характеризуется вполне явной правосторонней асимметрией и островершинностью. Отличие асимметрией от нормального распределения можно объяснить наличием большого числа публикаций по

прикладным аспектам **МТОС,** что прямо влияло на их объемные характеристики. **ТТГ(X)**-*распределение* можно разложить на два распределения **G(X**) – *основное* и **S(X) –** *дополнительное*, влияющее на поведение общего распределения **ТТГ(X),** в основном, в районе модально-медианного интервала (рис. **10, b**). Составляющая **S(X)** характеризует появление островершинности **ТТГ(X)**-распределения и может быть проинтерпретирована как большая интенсивность публикаций на уровне *кратких сообщений*, для которых в Трудах АН ЭССР был специальный раздел, а также небольших статей, посвященных отдельным вопросам **МТОС** и оперативно отражающих важнейшие результаты **ТТГ** в периоды повышенной международной научной активности в данном направлении.

Таким образом, ***кривые распределения*** широко используются для характеристики структуры изучаемого явления. Сопоставляя и анализируя их, получаем хорошую возможность судить об устойчивости процесса и степени влияния изменений, вносимых в процесс, на вариацию исследуемого признака. Особую роль при анализе *кривых распределения* играет *компьютерный* подход, позволяющий на основе известных пакетов удобно визуализировать весь процесс анализа и решать много важных задач, связанных с распределениями.

## 7.4. Элементы теории выборочного метода

До сих пор нами рассматривались *генеральные выборки*, т.е. *совокупности* в целом. ***Выборочным*** называют такое наблюдение, когда статистическому обследованию подлежат не все единицы исследуемой совокупности, а отобранные лишь определенным способом. При этом, цель такого статистического наблюдения состоит в том, чтобы по его *полученным* характеристикам судить о соответствующих характеристиках всей совокупности в целом. Поставленной цели оно достигает только при условии соблюдения определенной методологии отбора единиц совокупности, исключающей влияние различных субъективных и тенденциозных факторов. *Выборочное наблюдение* (**ВН**) включает следующие основные этапы:

1. *Постановка цели и основных задач выборочного наблюдения*
2. *Составление программы выборочного наблюдения, его плана и разработки статистических данных*
3. *Решение основных организационных вопросов выборочного наблюдения*
4. *Определение объема и методики отбора единиц исследуемой совокупности статистических данных*
5. *Проведение собственно выборочного наблюдения и регистрация соответствующих признаков*
6. *Обобщение результатов наблюдения, вычисление необходимых выборочных характеристик и ошибок выборки*
7. *Пересчет полученных выборочных характеристик на генеральную совокупность статистических данных*.

Этапы **1 - 3** и **5** *выборочного наблюдения* (**ВН**) осуществляются аналогично случаю сплошного статистического наблюдения и рассматривались нами выше. Остальные этапы будут нами обсуждаться с различной степенью полноты в настоящем разделе. **ВН** используется с целью получения экономии времени, людских и материальных ресурсов. В ряде случаев **ВН** просто является *единственно* возможным *видом* статистического *наблюдения*. Иногда **ВН** используется для уточнения результатов сплошного наблюдения. Если **ВН** проведено с соблюдением требуемой методики, то его данные верно характеризуют генеральную совокупность и практика его применения подтверждает этот вывод. Однако, перенесение результатов выборки на генеральную совокупность связано с ошибкой *репрезентативности* и даже при правильно проведенном отборе статистических единиц выборка не может в точности

представлять генеральную совокупность. Возникающие при этом ошибки случайны и их можно достаточно достоверно оценивать.

В основе теоретического обоснования *выборочного метода* лежит целый ряд предельных теорем теории вероятностей, иногда объединяемых под общим собирательным названием *закона больших чисел* (**ЗБЧ**). В социально-экономической статистике **ЗБЧ** – общий принцип, в силу которого количественные закономерности, присущие *массовым общественным* явлениям, четко проявляются только при достаточно большом количестве наблюдений. Тогда как в теории вероятностей **ЗБЧ** – ряд предельных теорем о математическом ожидании случайной величины (*переменной*). В теоретическом обосновании выборочного метода большую роль сыграли работы многих представителей русской и советской школ математической теории вероятностей П. Л. Чебышева, А.А. Маркова, А.М. Ляпунова, А.Я. Хинчина, А. Колмогорова и целого ряда других исследователей.

Предельные теоремы позволяют оценивать разность между генеральными и выборочными характеристиками совокупностей, отражающих поведение массовых случайных явлений. Поскольку изучаемые статистикой массовые явления тесно связаны с большим числом случайных влияний на них, то в ней используется основной вывод предельных теорем – *совокупное влияние большого числа случайных факторов при определенных условиях приводит к результату, практически, не зависящему от случая*. Так как **ВН** связано со случайными отклонениями выборочных характеристик от генеральных, то данный основной вывод позволяет утверждать, что результаты выборки достоверны, если она достаточно большого объема, т.е. при достаточно большом объеме выборки ее характеристики довольно хорошо соответствуют генеральным характеристикам.

Предельные теоремы исходят из нормального закона распределения случайных величин, из которого, в частности, следует, что большая часть величин группируется около генеральной средней (*рассмотренное выше правило трех сигм*). *Нормальное распределение* позволяет оценивать частоту появления ошибок для данного размера средней. Следствия из предельных теорем позволяют абстрагироваться от ошибок регистрации при выборке и ограничиться только собственно ей присущими ошибками. Наиболее обобщающей предельной теоремой, относящейся к выборочному методу, является теорема Чебышева-Ляпунова о расхождении значений *выборочной* и *генеральной* средних величин (*эти расхождения называются ошибками выборки*). ***Ошибки выборки*** являются как следствием погрешностей *регистрации* наблюдения, так и репрезентативности, свойственной только самой выборке.

Теорема Чебышева-Ляпунова устанавливает, что *ошибка репрезентативности* при достаточно большом объеме (**n**) выборки будет сколь угодно малой, а точнее имеет место следующее определяющее соотношение:

$$P(|\overline{X}_s - \overline{X}_g|) > 1 - \frac{1}{t^2} = P(t) \qquad (53)$$

где $|\overline{X}_s - \overline{X}_g|$ – абсолютная величина расхождения между *выборочной* и *генеральной* средними (*ошибка репрезентативности*), **d** – среднее квадратическое отклонение выборочной средней от генеральной [*средняя ошибка выборки*; *она зависит от колеблемости признака* ($\sigma_g$) *в генеральной совокупности и объема* (**n**) *самой выборки*, *т.е.* $d = \sigma_g/\sqrt{n}$, *а также от способа выборки*]. Из данного весьма важного результата (53) можно сделать вывод, что с вероятностью не меньшей, чем **P(t),** *репрезентативная ошибка* не превысит величины **t*d,** т.е. имеет место соотношение:

$$|\overline{X}_s - \overline{X}_g| \le t*d = \frac{\sigma_g}{\sqrt{n}}$$

Например, для **t = 3** вероятность справедливости неравенства $|\overline{X}_s - \overline{X}_g| \le 3*d$ превышает значение **P(3)** = 0.89, тогда как для **t = 4** она уже достигает значения **P(4)** = 0.94. С другой стороны, *выборочная* и *генеральная* дисперсии связаны соотношением $\sigma_g^2 = \sigma_s^2 * n/(n-1)$, т.е. в случае достаточно больших значений **n** мы можем использовать следующее асимптотическое соотношение $\sigma_g^2 \approx \sigma_s^2$. Таким образом, для вычисления *генеральной* средней по результатам *выборки* мы получаем следующее определяющее соотношение:

$$P(|\overline{X}_s - \overline{X}_g| \le t*\frac{\sigma_s}{\sqrt{n}}) > 1 - \frac{1}{t^2} = P(t)$$

из которого следует, что с вероятностью **P(t)** значение генеральной средней $\overline{X}_g$ находится в интервале $I_s = (X_s - t*\sigma_s/\sqrt{n},\ X_s + t*\sigma_s/\sqrt{n})$. Величину **O(t) = t*d** называют *предельной ошибкой выборки*, из которой следует, что с увеличением значения величины **t** растет и вероятность принадлежности средней генеральной указанному *доверительному* **Ic**-интервалу, но, вместе с тем, растет и сам доверительный интервал (*величина предельной ошибки*). Следует иметь в виду, что ошибка **O(t)** выборки вычисляется по-разному для различных способов отбора единиц из генеральной совокупности.

А.М. Ляпунов доказал, что вероятность **P(t)** отклонений выборочной средней от генеральной при достаточно больших объемах (**n**) производимой выборки подчиняется нормальному закону распределения, усилив тем самым первоначальный результат Чебышева. Поэтому, на основе правила трех сигм (**t = 3**) получаем соотношение для предельной ошибки выборки **O(3) = 3*$\sigma_s/\sqrt{n}$,** которое справедливо с вероятностью **P(3) = 0.997.** Предположим теперь, что выборка объемом **n=100** дала для величин $\overline{X}_s$ и $\sigma_s$ некоторого признака соответственно значения 19.42 и 0.58. Тогда с вероятностью **P(3)=0.997** значение средней $\overline{X}_g$ генеральной совокупности лежит в следующем *доверительном* интервале: **Is** = (19.42 - 3*0.58/10, 19.42 + 3*0.58/10) = (19.246, 19.594), т.е. локализуется достаточно точно; при этом, $\sigma_g = \sigma_s * 1.005 = 0.583$, что, практически, совпадает с *априорным* выборочным средне квадратическим отклонением.

Для оценки генеральной частости (*вероятности*) некоторого признака по ее выборочной используется теорема Бернулли, гласящая, что: *при достаточно большом объеме* (**n**) *выборки с вероятностью, сколь угодно близкой к* **1,** *можно утверждать, что выборочная* (**w**) *частость признака сколь угодно мало отличается от соответствующей ей генеральной* (**p**) *частости* (*вероятности*):

$$P(t) = \lim_{n \to \infty} P(|w-p| \le t*d) = 1 \qquad (54)$$

Так как величина **P(t)** следует нормальному закону распределения, то ее значение легко вычислять по соответствующим статистическим таблицам или интегральной формуле в зависимости от значения доверительного уровня **t.** Пусть, например, **P(3) = 0.997.** Тогда с вероятностью **P = 0.997** значение величины **p** будет лежать в интервале **If = (w -3*$\sigma_g/\sqrt{n}$, w+3*$\sigma_g/\sqrt{n}$),** что в практических расчетах вполне приемлемый уровень достоверности. Так как величина $\sigma_g$ среднего квадратического отклонения в генеральной совокупности для альтернативного признака равна **p*q (q = 1 - p)**, как было показано в разделе **2.4,** то для нашего случая имеем $d = \sqrt{p*q/n}$. Но так как величина **p*q** нам неизвестна, то мы вынуждены принять ее равной величине **w*(1 - w),** что приводит к следующей формуле для *предельной ошибки* выборочной частости, а именно: $O(t) = t*\sqrt{w*(1-w)/n}$. Это, в свою очередь, полностью дефинирует доверительный интервал для значения генеральной частости (*вероятности*) **p** с

вероятностью **P(t)** как $\mathbf{If} = (w - t*\sqrt{w*(1-w)/n},\ w + t*\sqrt{w*(1-w)/n})$. Предположим, что для выборки объемом **n = 100** получено значение частости **w =** 0.58 некоторого признака. Тогда с учетом сказанного выше легко вычисляем доверительный интервал вида

$$\mathbf{If} = (0.58 - 3*\sqrt{0.58*0.42/100},\ 0.58 + 3*\sqrt{0.58*0.42/100}) = (0.432, 0.728)$$

в котором с вероятностью **P(3) = 0.997** и будет находится значение генеральной частости исследуемого признака.

В целом ряде статистических исследований приходится иметь дело с так называемыми *малыми выборками* (**MB**), когда выборочная совокупность содержит небольшое (*обычно до* **30**) число единиц из генеральной совокупности. В этом случае *среднее квадратическое отклонение* вычисляется по следующей простой формуле: $\sigma_{ss} = \sqrt{n*(n-1)}*\sigma$, т.е. относительно обычной выборки вводится некоторый поправочный коэффициент, определяемый только объемом выборки. Вероятностная оценка результатов малой выборки отличается от обычной тем, что средняя выборочная ошибка подчиняется не нормальному закону распределения, а распределению Стьюдента (*раздел* **2.5**), *предельным* для которого служит *первое* распределение (*удовлетворительное совпадение наблюдается уже при значении* **n = 30**). Поэтому для случая **MB** вычисляется не вероятность **P(t),** а вероятность **P(t, n),** зависящая от доверительного **t**-уровня и объема **n** выборки.

Таким образом, *доверительный интервал* $\mathbf{I_a} = [X_{ss} - O(t),\ X_{ss} + O(t)]$, содержащий само значение генеральной средней, определяется с вероятностью **P(t, n),** которая легко вычисляется по соответствующей статистической таблице либо по интегральной формуле. Предположим, что по малой выборке **SS** объемом **n = 5** получены значения ее средней $\overline{X}_{ss} = 19.47$ и средне квадратического отклонения $\sigma_{SS} = 0.47$. Тогда по табл. вероятностей Стьюдента [91, 125] определяем, что с вероятностью **P(3, 5) = 0.96** величина *генеральной средней* будет находиться в следующем *доверительном интервале* **Ia** = (19.47 - 3*0.53/√5, 19.47 + 3*0.53/√5) = (18.759, 20.181). Таким образом, вычисления доверительных интервалов для генеральной средней в случае **MB** отличается от случая обычной выборки только методом вычисления предельной ошибки выборочной средней и вероятностью доверительного **Ic**-интервала.

Для иллюстрации вышерассмотренного материала мы дадим комплексный пример оценки *генеральной средней* на основе результатов анализа *обычных* и *малых* выборок (рис. **11**). С этой целью в качестве *генеральной совокупности* мы выбираем некоторую абстрактную *совокупность* цитирования публикаций **ТТГ** по **МТОС** в целом, в то время как в качестве *обычной* выборки из совокупности мы определяем статистические данные, относящиеся к цитированию наших публикаций, представленных в табл. **9** (*графа* **5**).

```
> restart;  with(SimpleStat):  Digits:= 6:  Or := (t, L) -> evalf(t*sqrt(Ds(L))/sqrt(nops(L)));
```

$$Or := (t, L) \rightarrow \mathrm{evalf}\left(\frac{t\sqrt{SimpleStat:-Ds(L)}}{\sqrt{\mathrm{nops}(L)}}\right)$$

```
> cnorm := t -> evalf((1/sqrt(2*Pi))*int(exp(-x^2/2), x = -infinity .. t));
```

$$cnorm := t \rightarrow \mathrm{evalf}\left(\frac{1}{\sqrt{2\,\pi}}\int_{-\infty}^{t} e^{\left(-\frac{x^2}{2}\right)}\,dx\right)$$

```
> P := (t, m) -> evalf(1/sqrt(Pi*m)*GAMMA((m+1)/2)/GAMMA(m/2)*int((1+x^2/m)^(-(m+1)/2),
      x= -infinity .. t));
```

$$P := (t, m) \rightarrow \text{evalf}\left( \frac{\Gamma\left(\frac{1}{2}m + \frac{1}{2}\right)}{\sqrt{\pi\, m}\,\Gamma\left(\frac{1}{2}m\right)} \int_{-\infty}^{t} \left(1 + \frac{x^2}{m}\right)^{\left(-\frac{1}{2}m - \frac{1}{2}\right)} dx \right)$$

```
> I1:= (L, t) -> SR(L) - Or(t, L);  I2:=(L, t) -> SR(L) + Or(t, L);  A:= [0, 33, 38, 42, 76, 94, 96, 98, 102,
  125, 145, 174, 186, 205, 220, 222, 236, 240, 266, 284, 292, 304, 314, 316, 320, 336, 345, 384, 395, 405]:
> `Average for A-sample` = evalf(SR(A));
```

*Average for **A**-sample* = 209.767

```
> `Variation for A-sample` = evalf(Ds(A));
```

*Variation for **A**-sample* = 13732.3

```
> `Confidence interval I[A]` = [I1(A, 3), I2(A, 3)];
```

*Confidence interval I[A]* = [145.582, 273.952]

```
> `Probability of hit of general average into I[A]` = cnorm(3);
```

*Probability of hit of general average into **I[A]*** = 0.998645

```
> G:= rand(1 .. nops(A)):  SS := [seq(A[G()], k=1 .. 20)]:  SS;
```

[304, 292, 384, 316, 236, 102, 94, 102, 33, 284, 336, 42, 384, 292, 220, 336, 96, 304, 384, 405]

```
> `Average for SS-sample` = evalf(SR(SS));
```

*Average for **SS**-sample* = 247.300

```
> `Variation for SS-sample` = evalf(Ds(SS));
```

*Variation for **SS**-sample* = 14536.0

```
> `Confidence interval I[SS] for general average` = [I1(SS, 3), I2(SS, 3)];
```

*Confidence interval I[SS] for general average* = [166.422, 328.178]

```
> `Probability of hit of general average into I[SS]` = P(3, nops(SS));
```

*Probability of hit of general average into **I[SS]*** = 0.994247

**Рис. 11. Вычисление *генеральной средней* совокупности на основе *обычной* и *малой* выборок**

Для оценки *средней* генеральной совокупности на основе *обычных* и *малых* выборок из нее используется вышеупомянутый пакет *Maple*. В *первой* части *Maple*-программы определяются функции для вычисления *средней* (**SR**), *дисперсии* (**Ds**), *средне квадратичного отклонения* (**σ**) и *предельной ошибки* (**Or**) для произвольной выборки. При этом, первые две процедуры **SR** и **Ds** находятся в нашей библиотеке в программном модуле **SimpleStat,** описанном в книге [302, 303]. Затем определяются процедуры для плотности нормального распределения (***cnorm***) и интеграла Стьюдента {**P(t, m)**}. На основе данных формул затем будут прямо вычислены вероятности попадания средних в соответствующие им доверительные интервалы. Наконец, определяются общие формулы для вычисления *доверительных интервалов* [***I1***, ***I2***] для средней.

Во *второй* части программы на основе данных из табл. **9** (*графа* **5**) определяется **A**-список, соответствующий обычной выборке, и на основе вышеупомянутых формул вычисляются такие показатели как *средняя*, *дисперсия* и *доверительный интервал* для *общей* средней, а также *вероятность* попадания *общей* средней в данный интервал. Вычисленные показатели имеют следующие значения, а именно: **SR(A)** = 210, **Ds(A)** = 13732.2, [**I1(A, 3)**, **I2(A, 3)**] = [146, 274] и

**P(3) = *cnorm*(3)** = 0.9986. При этом, значения для **SR** и границ доверительного интервала округлены до целых чисел согласно здравому смыслу (*количество цитирований может быть только целым числом*). Таким образом, для *доверительного уровня* **t = 3** *общая* средняя попадет в *доверительный* интервал [146, 274] с вероятностью **P(3)=0.9986.** При вычислении вероятности использовалась непосредственно интегральная формула нормального распределения.

Наконец, в *третьей* части программы посредством стохастической процедуры, *базирующейся* на использовании встроенного генератора псевдослучайных целых чисел (*процедура* ***rand***), генерируется малая выборка **SS** из 20 единиц обычной **A**-выборки, созданной выше. **SS** предназначена для оценки средней *генеральной* совокупности и *обычной* **A**-выборки. Для вычисления вероятности попадания средней в доверительный интервал непосредственно используется интегральная формула Стьюдента **P(t, m)**. Проведенные нами вычисления показывают, что средние генеральной совокупности и обычной выборки с вероятностью **P(3, 20) = 0.9942** принадлежат доверительному интервалу [166, 328], который намного шире интервала, полученного ранее на основе *обычной* **A**-выборки. Читателю рекомендуется более детально рассмотреть данный пример, который представляет также и самостоятельный практический интерес (рис. **11**). При этом, для полного понимания вышеприведенной *Maple*-программы читатель может ограничиться информацией по пакету, например, в рамках книг [99, 139, 140-143, 158].

Однако из всего сказанного следует, что вместо *обычной* выборки (*скажем*, *в целях экономии*) невозможно повсеместно применять **МВ**. Выводы на ее основе действительны лишь тогда, когда распределение исследуемого признака в *генеральной* совокупности является *нормальным* или близким к нему; при этом, точность результатов при **МВ** ниже (*в целом ряде случаев существенно*), чем при выборках большого объема. Поэтому **МВ** в социально-экономических исследованиях следует применять с большой осторожностью. Более широкое применение они находят в технических приложениях и медико-биологических исследованиях, где на их основе производится изучение взаимосвязи между признаками посредством дисперсионного анализа.

При планировании выборочного наблюдения возникает вопрос о необходимом *объеме* (**n**) *выборки*, который вычисляется, исходя из формулы *предельной выборочной ошибки*, как $\mathbf{n = (t*\sigma)^2/d^2}$. Данная формула показывает, что с ростом *средней выборочной ошибки* (**d**) весьма существенно уменьшается *необходимый объем* (**n**) выборки. Конкретная формула на предмет вычисления величины **n** выводится из формулы *предельной выборочной ошибки*, существенно зависящей от методов отбора единиц из *генеральной* совокупности.Так, например, для метода *случайной повторной выборки*, являющейся наиболее простой, имеет силу все вышесказанное. Другие методы и способы отбора имеют свои особенности при проведении выборки и свои расчетные формулы для вычисления средней ошибки (**d**) выборки. Подробнее с подобными вопросами можно ознакомиться, например, в книгах [29, 40,43,58,62,70,91,128,155,182,188,256]. Отечественная статистика накопила достаточно богатый практический опыт применения выборочного метода, из которого следует отметить такие основные его приложения как: *выборочная разработка данных*; обследования *бюджетные*, *семей рабочих*, *колхозников* и *служащих*; *перепись населения*; *контроль качества продукции* и целый ряд других.

## 7.5. Проверка статистических гипотез

***Общие соображения.*** Рассматривая в предыдущем разделе **7.4** элементы теории выборочного теории метода, мы отметили, что *выборочные* характеристики являются *оценками* параметров генеральной совокупности, которые, как правило, неизвестны. В этом отношении были даны некоторые важные оценки. Ниже мы рассмотрим сравнительные оценки общих параметров

на основе *различий*, наблюдаемых между сравниваемыми выборками. Данное обстоятельство чрезвычайно важно, ибо, фактически, любое статистическое исследование не обходится без сравнения. Преимущества сравниваемых выборок определяются, как правило, различиями между средними и другими выборочными показателями, которые являются случайными переменными, сопровождаемыми ошибками репрезентативности. В этом случае проблема уверенности в выборочной природе различия с его ошибкой решается на основе проверки гипотезы (*статистического теста*), т. е. *гипотезы* или *предположения* относительно *параметров сравниваемых* выборок, которое выражается в терминах вероятности и может быть проверено на основе *выборочных* характеристик. Логика, используемая в статистических тестах, которые мы будем обсуждать ниже, подобна логике, используемой во многих судах, когда *обвиняемый* полагается *невинным*, и обвинитель должен доказать обратное, снимая все разумные *сомнения*. Для наших же целей этому служит статистическая проверка гипотез.

***Статистическая проверка гипотез*** – система приемов в статистике, предназначенная для проверки соответствия опытных данных проверяемой гипотезе. К проблеме статистической проверки гипотез приводит большое число связанных с практикой вопросов, возникающих в приложениях, например, сравнение урожайности сортов каких-либо сельскохозяйственных культур, эффективности лекарственных препаратов и др. *Правило*, по которому принимается или отклоняется данная гипотеза, называют ***статистическим критерием***. ***Статистическая гипотеза*** представляет собой некоторое предположение о законе распределения случайной величины или о параметрах этого закона, формулируемое на основе выборки. Примерами статистических гипотез являются предположения: генеральная совокупность распределена по нормальному закону; математические ожидания двух экспоненциально распределенных выборок равны друг другу. В первой из них высказано предположение о виде самого закона распределения, а во второй – о параметрах двух распределений. Гипотезы, в основе которых нет допущений о конкретном виде закона распределения, называют ***непараметрическими,*** в противном случае – ***параметрическими***.

Одной из важнейших задач профессионального статистика является проверка выдвинутых им же *предположений* или *гипотез*. Чем же отличаются *статистические гипотезы* от обычных предположений? Прежде всего, тем, что статистических гипотез всегда две и они являются взаимоисключающими. Одна из них (*обычно та, которую предполагают отклонить*) называется ***нулевой гипотезой*** (**Но**), вторая – ***альтернативной гипотезой*** (**Н1**), отрицающей *нулевую*.

Вся проблема заключается именно в нулевой гипотезе – её надо построить, сформулировать так, чтобы иметь возможность найти интересующие нас вероятности в условиях истинности этой гипотезы. С процедурами проверки статистических гипотез неразрывно связано еще одно, непривычное для обычных расчетных работ, понятие ***уровня значимости*** результатов наблюдений. Будем далее использовать 5% уровень значимости, как это принято почти во всех прикладных направлениях статистики, в том числе и в экономике.

Предположение, что нет никакой связи, обычно называют ***нулевой гипотезой,*** так как она аннулирует наш прогноз. Результат, который мы желаем продемонстрировать, называют ***альтернативной гипотезой*** или ***исследовательской гипотезой.*** В классической постановке сущность нулевой гипотезы состоит в предположении, что различия между генеральными параметрами сравниваемых выборок равняются *нулю*, и несоответствия, наблюдаемые между *выборочными* характеристиками, вызваны не *систематическими*, а исключительно *случайными* причинами. Так, например, если некоторая выборка извлечена из *нормально* распределенной совокупности с параметрами $\overline{\mathbf{X}}\mathbf{1}$ и $\boldsymbol{\sigma}\mathbf{1}$, тогда как другая выборка извлечена из совокупности с параметрами $\overline{\mathbf{X}}\mathbf{2}$ и $\boldsymbol{\sigma}\mathbf{2}$, то ***нулевая гипотеза*** следует из следующего предположения $\overline{\mathbf{X}}\mathbf{1} - \overline{\mathbf{X}}\mathbf{2} = \mathbf{0}$ и $\boldsymbol{\sigma}\mathbf{1} - \boldsymbol{\sigma}\mathbf{2} = \mathbf{0}$. Тогда как *альтернативная* гипотеза следует из предположения: $\overline{\mathbf{X}}\mathbf{1} \neq \overline{\mathbf{X}}\mathbf{2}$ и $\boldsymbol{\sigma}\mathbf{1} \neq \boldsymbol{\sigma}\mathbf{2}$.

Для проверки принятой гипотезы (*и следовательно уверенности в оценке генеральных параметров на основе выборочных данных*) используются *величины*, *функции распределения* которых известны. Данные величины, называемые ***критериями достоверности,*** позволяют выявлять в каждом конкретном случае факт соответствия *выборочных* показателей принятой гипотезе. Функции распределения данных величин представлены специальными таблицами, которые содержат значения функции для различных *чисел степеней свободы* **k** или *размера выборки* **n** наряду с *уровнем значимости* (**α)**.

***Уровень значимости*** или вероятность ошибки, допустимой при оценке принятой гипотезы могут отличаться. Обычно, при проверке статистических гипотез рассматриваются три уровня значимости, а именно: **5%** (*вероятность* **P** *ошибочной оценки* **P=0.05**), **1%** (**P=0.01**) и **0.1%** (**P=0.001**). В экономических исследованиях, социальных и во многих других исследованиях вполне достаточно ограничиться уровнем значимости в **5%.** Кроме того, нулевая гипотеза не отклоняется, если в результате исследования будет определено, что вероятность *погрешности* оценки относительно точности принятой гипотезы превышает **5%,** т. е. **P> 0.05.** Если **P <0.05,** то принятая гипотеза должна быть отклонена на принятом уровне значимости (**α**). Ошибка в этом случае возможна не более, чем в **5%** случаев, что достаточно маловероятно. При более ответственном исследовании *уровень значимости* может быть *понижен* до **1%** или даже до **0.1%.**

Проверка статистических гипотез предполагает выполнение следующих *пяти* этапов:

1. Исследователь формулирует исходное утверждение, подлежащее эмпирической проверке. Это утверждение базируется на *предыдущем* опыте (*результатах предшествующих эмпирических исследований*, *теории или догадке*) и называется *нулевой гипотезой* (**Ho**). Формулируется также противоположное утверждение — альтернативная гипотеза (**H1**). В ходе проверки гипотезы мы должны будем принять решение о том, какое из утверждений является верным в свете полученных эмпирических данных.
2. Принимается вероятность ошибочного отвержения *нулевой гипотезы* — *уровень значимости* (**α**). Как правило для **α** принимаются значения 0.05, 0.01 или 0.001.
3. Извлекается выборка и для полученных эмпирических данных определяется значение статистического критерия и вероятность его получения для ситуации, когда *нулевая гипотеза* верна.
4. В случае, если *вероятность* получения критерия оказывается *меньше* установленного *уровня значимости*, исследователь отвергает нулевую гипотезу. В противном случае мы говорим, что полученные эмпирические данные не позволяют отвергнуть нулевую гипотезу, т.е. гипотеза принимается с выбранныи уровнем значимости.
5. Рассматриваются следствия принятого решения — полученные результаты подвергаются интерпретации.

На практике часто рассматриваются проверки гипотез о равенстве *дисперсий* и *средних* в двух генеральных совокупностях по сделанным выборкам из обоих совокупностей.

***Параметрические и непараметрические статистические критерии.*** В областях социальных, экономических и целого ряда других исследований применяются два вида статистических критериев – *параметрические* и *непараметрические*. ***Параметрические критерии*** строятся на основе параметров заданной совокупности (*например*, *средней* ($\overline{\mathbf{X}}_\mathbf{s}$) *и среднего квадратичного отклонения*) и представляют собой функции этих параметров. Тогда как ***непараметрические критерии*** (*тесты*) представляют собой функции, зависящие непосредственно от *значений* и их *частот* изучаемой переменной заданной совокупности. Необходимо иметь в виду, что если *параметрические* критерии служат для проверки гипотезы о параметрах совокупностей, распределенных согласно нормальному закону, то *непараметрические* служат для проверки

рабочих гипотез независимо от формы распределения совокупностей, из которых выбраны сравниваемые выборки.

В последние годы большую популярность приобрели именно *непараметрические* критерии (*Манна-Уитни*, *Ван дер Ваардена*, *Уилкоксона и др.*). Их достоинством является то, что они не содержат ограничений, вытекающих из гипотез о типе *распределения* случайных величин, а опираются на единый принцип — *непрерывности* распределений. Эти критерии применимы и для анализа *порядковых данных*. Однако, по сравнению с *параметрическими* методами они менее чувствительны к различиям в выборках. Чаще всего непараметрические критерии используются для сравнения эмпирического распределения с теоретическим, в частности, при проверке имеющейся статистической совокупности на принадлежность к нормальному типу распределений.

Особое преимущество *непараметрических* методов по сравнению с *параметрическими* состоит в том, что они являются довольно легкими для понимания. Для полного понимания этих методов (*в большинстве своем*) необходимо только самое элементарное знание математики. Главное же неудобство *непараметрических* методов состоит в том, что они не совсем пригодны к сложным статистическим исследованиям, в которых исследуется достаточно большое число переменных, тогда как *параметрические* методы типа *дисперсионного* анализа хорошо отвечают таким ситуациям.

В случае нормального распределения переменной *параметрические* методы имеют большую потенцию, чем *непараметрические*. Они способны более уверенно отвергать *нулевую гипотезу*, если она ошибочна. Поэтому, во всех случаях, когда сравниваемые выборки извлечены из нормально распределенных совокупностей, следует предпочесть параметрические методы. В случае довольно больших отличий распределений переменной от нормального необходимо применять *непараметрические* критерии, которые в такой ситуации являются более *мощными*. Тогда как в ситуациях, когда *варьирующие* переменные выражаются *не числами*, а некоторыми условными знаками, возможно использование только *непараметрических* критериев.

***Параметрические критерии*** (*тесты*). *Из параметрических* критериев в общей статистике и большинстве отраслевых статистик весьма широко используются **t**-критерий Стьюдента и **F**-критерий Р. Фишера. Первый критерий используется для сравнительной оценки средних, в тогда как второй критерий используется для оценки дисперсии. Рассмотрим несколько более детально оба критерия как типичные представители *параметрической* статистики.

**t-критерий Стьюдента** (**t-*распределение***). Использованию формулы Гаусса-Лапласа, которая определяет нормальный закон распределения (26), для сравнительной оценки средних чисел препятствует то обстоятельство, что в качестве аргументов эта формула включает основные параметры — *среднюю* и *дисперсию* (*которые*, *как правило*, *неизвестны*), тогда как при обработке и сравнении выборок необходимо использовать не *генеральные* характеристики, а *выборочные*. Поэтому английский математик Госсет (*псевдоним Стьюдент*) в 1908 ввел закон *распределения* следующей величины t:

$$\mathbf{t = \frac{\overline{X}_s - \overline{X}}{\sigma_s/\sqrt{n}}} \qquad (54.1)$$

где $\overline{\mathbf{X}}_\mathbf{s}$ и $\overline{\mathbf{X}}$ — выборочная и генеральная средние соответственно; $\boldsymbol{\sigma}_\mathbf{s}$ — выборочное среднее квадратичное отклонение, и **n** — размер выборки. Было установлено, что *отношение* **t** *разницы* между выборочной и генеральной средними к ошибке выборочной средней распределено непрерывно согласно следующей формуле:

$$\mathbf{f(t) = C*(1+\frac{t^2}{n-1})^{1-n/2}} \qquad (54.2)$$

где **C** — константа, зависящая только от степеней свободы **k = n - 1.**

Открытый Стьюдентом и теоретически обоснованный Р. Фишером, закон **t**-распределения служит основой так называемой ***теории малых выборок,*** характеризующей распределение выборочных средних в нормально распределенных совокупностях в зависимости от размера выборки. **t**-распределение зависит только от числа степеней свободы **k=n - 1** и с увеличением *размера* выборки **n, t**-*распределение* быстро сходится к *стандартному нормальному распределению* с параметрами $\overline{X}$ и **σ = 1** и уже для **n ≥ 30** не отличается от него.

Для выборок, чей размер **n** превышает **30,** величтна **t** распределяется нормально и не зависит от количества наблюдений. Тогда как в случае **n < 30,** характер **t-**распределения существенно зависит от количества наблюдений **n.** Для практического *использования* t-распределения была составлена специальная таблица, имеющаяся во многих статистических руководствах (*см., например, табл.* **V** [155]). Данная таблица содержит критические точки для различных *уровней значимости* (**α**) и *чисел степеней свободы* **k = n - 1.** Принцип использования данной таблицы рассматривается на примере ниже. **t**-*распределение* может использоваться для многих важных прикладных задач таких как: оценки разности средних двух независимых выборок, оценка разности между выборочными и генеральными средними и т.д. Здесь мы ограничим себя рассмотрением примера, относящегося к первой задаче оценки разности средних.

***Оценка разности средних двух независимых выборок.*** Сравнивая две независимые выборки, выбранные из нормально распределенных совокупностей со средними **h1** и **h2,** мы можем предположить, что **h1 - h2 = D** и дисперсия этой разности равна $\sigma D^2$. Значения генеральных параметров неизвестны; однако, весьма просто найти значения выборочных средних **s1** и **s2,** и разность между ними, а именно: **s1 - s2 = d.** Тогда нулевая гипотеза (**Ho**-*гипотеза*) состоит в предположении о наличии равенства **h1 = h2.** При этом, в качестве критерия проверки **Ho**-гипотезы служит следующее отношение:

$$t = [(s1 - s2) - (h1 - h2)]/S_d$$

где **t** — переменная, соответствующая **t-**распределению (54.2) с числом степеней свободы **k = (n1 - 1) + (n2 - 1) = n1 + n2 - 2, Sd** — ошибка указанной разности. Так как согласно **Ho**-гипотезе **h1 - h2 = 0,** то **t**-критерий выражается посредством отношения разности выборочных средних к ошибке, а именно:

$$t = (s1 - s2)/S_d = d/S_d$$

**Ho**-гипотеза отвергается, если фактически вычисленное значение (*обозначаемое* **tf**) **t-**критерия превысит или будет равно критическому значению **tst** этой величины для принятого *уровня значимости* (**α)** и *числа степеней свободы* **k=n1+n2-2,** т.е. при условии **tf ≥ tst.** При этом, ошибка разностия средних рассчитывается согласно следующим формулам:

$$\text{if } n1 = n2 = n: \quad S_d = \sqrt{\frac{\sum_j (x_j - s1)^2 + \sum_j (x_j - s2)^2}{n*(n-1)}} \tag{54.3}$$

$$\text{if } n1 \neq n2: \quad S_d = \sqrt{\frac{\sum_j (x_j - s1)^2 + \sum_j (x_j - s2)^2}{n1 + n2 - 2} * \left(\frac{n1 + n2}{n1*n2}\right)} \tag{54.4}$$

На основе данного алгоритма просто вычислять значение **tf** и число степеней свободы **k** для двух произвольных выборок, выбранных из нормально распределенных совокупностей, и на основе специальной таблицы критических точек **t**-критерия Стьюдента мы можем оценить

достоверность **Ho**-гипотезы относительно равенства средних генеральных совокупностей. Для упрощения вычислений на основе данного алгоритмав среде *Maple* была реализована процедура ***T_test_AV*, возвращающая список вида **[k, tf].** В разделе **10.5.3** представлены как описание процедуры, так и пример ее применения для анализа двух конкретных выборок.

При использовании **t**-критерия необходимо иметь в виду, что принятие **Ho**-гипотезы нельзя рассматривать в качестве *строгого* доказательства *равенства* между *неизвестными* параметрами совокупностей, из которых были извлечены сравниваемые выборки. Бывает, при повторных выборках **Ho**-гипотеза, возможно, может оказаться необоснованной. Более того, даже в случае отказа от **Ho**-гипотезы мы не должны спешить с окончательными выводами. Необходимо отметить, что вышеописанное применение **t**-критерия предполагает равенство разностей сравниваемых выборок. Иначе, t-величина критерия (*теста*) рассчитывается по несколько иной формуле, с которой можно ознакомиться, например, в книге [155] {*формулы* (75)}.

Корректное применение **t**-критерия предполагает *нормальное* распределение совокупностей, из которых извлекаются сравниваемые выборки наряду с равенством генеральных *дисперсий*. Если эти условия не выполняются, то использовать **t**-критерий не следует. В таких случаях более эффективными будут *непараметрические* критерии, которые рассматриваюся ниже.

**F-критерий Фишера (F-*распределение*).** Для проверки **Ho**-гипотезы о равенстве генеральных дисперсий нормально распределенных генеральных совокупностей **t**-критерий оказывается недостаточно точным, особенно при оценках разности дисперсий в случае *малых выборок*. В поисках лучшего критерия Р. Фишер установил, что вместо разности выборочных дисперсий $\sigma 1_s - \sigma 2_s$ удобнее использовать разность между натуральными логарифмами этих величин, а именно: : $\ln(\sigma 1_s) - \ln(\sigma 2_s)$, где $\sigma 1_s \geq \sigma 2_s$. Эта разность, определенная Фишером как **z**-величина, распределена *нормально* в случае статистических совокупностей как большого, так и среднего размера. При этом, при определении **z**-величины нормальные логарифмы можно заменить десятичными. При дальнейшем развитии данного метода Дж. Снедекор предложил вместо логарифмов отношений использовать отношения выборочных дисперсий, обозначив в честь Р. Фишера этот показатель буквой **F,** а именно:

$$F = \frac{\sigma 1_s^2}{\sigma 2_s^2} \qquad for \quad \sigma 1_s^2 \geq \sigma 2_s^2 \tag{54.5}$$

Согласно формуле (54.5), имеет место очевидное соотношение $F \geq 1$ и $F = 1$ в случае равенства выборочных дисперсий. Чем больше **F**-величина, тем более значительная разность между *выборочными дисперсиями*, и наоборот. Величина **F** имеет *непрерывную* функцию *распределения* и зависит только от чисел степеней свободы $k1 = n1 - 1$ и $k2 = n2 - 1$, где **n1** и **n2** — размеры обоих выборок соответственно.

Величина **F** полностью определена выборочными дисперсиями и не зависит от генеральных параметров при условии, что сравниваемые выборки, характеризуемые дисперсиями $\sigma 1_s$ и $\sigma 2_s$, отобраны из генеральных совокупностей с равными дисперсиями или из той же самой генеральной совокупности. Функция распределения возможных значений величины **F** для *малых выбо*рок имеет форму *асимметричной* кривой, которая с увеличением размеров *выборок* ($n \to \infty$) приближается к кривой *нормального* распределения. Функция **F-***распределения* сведена в таблицу для уровней значимости в **5%** (**P=0.05**) и **1%** (**P=0.01**) совместно с числами степеней свободы **k1** для большей дисперсии и **k2** для меньшей дисперсии.

Для практического использования **F-***распределения* составлена специальная таблица, которая имеется во многих статистических руководствах (*см*., *например*, табл. **VI** [155] *или* [223, 225]).

Данная таблица содержит критические точки для двух уровней значимости (**α = 5%** и **α = 1%**) совместно с числами степеней свободы **k1 = n1 - 1** и **k2 = n2 - 1.** Метод использования данной таблицы довольно прост и состоит в следующем. Если сравниваемые выборки извлечены из той же самой *генеральной* совокупности или из разных совокупностей с равными *дисперсиями*, то **Ff**-значение **F**-критерия не будет превышать значения **Fst** *критической точки*, соответствуя принятому *уровню значимости* (**α**) {**5%** *или* **1%**} и числа степеней свободы **k1** и **k2.** Если *выборки* были извлечены из разных генеральных совокупностей с разными дисперсиями, то $\mathbf{F_f \geq F_{st}}$ и **Ho**-гипотеза должна быть отвергнута.

На основе данного алгоритма весьма просто вычислять *значение* **Ff** и числа степеней свободы **k1** и **k2** для двух произвольных выборок из нормально распределенных совокупностей. И на основе специальной таблицы *критических точек* **Fst F**-критерия можно оценить *достоверность* **Ho**-гипотезе относительно равенства дисперсий генеральных совокупностей. Для упрощения *вычислений* на основе *данного* алгоритма в среде *Maple* была реализована процедура ***F_test_Ds*,** которая возвращает список вида **[k1, k2, Ff].** В разделе **10.5.3** представлены как описание, так и пример использования процедуры для анализа двух конкретных выборок. Таким образом, из последнего и предпоследнего примеров мы можем сделать следующий вывод: с высокой достоверностью вышеупомянутые малые выборки **L1** и **L2** (*раздел* **10.5.3**) извлечены нами из нормально распределенных совокупностей с идентичными *средними* и *дисперсиями* [302, 303].

В заключении рассмотрения *параметрических критериев* (*тестов*) еще раз следует отметить, что они дают весьма приличные результаты только для случайных выборок из нормально распределенных совокупностей. Иначе следует использовать непараметрические критерии, которые рассматриваются ниже. Имеется достаточно большое количество *параметрических* критериев, ориентируемых на различные применения, и читатель, интересующийся данной проблемой, может ознакомиться с ней в книгах [46, 48, 49, 53, 55, 91, 155, 157, 175, 182, 236, 238, 256, 260, 264, 265, 274, 276, 301, 305-311]. В частности, для интересующихся использованием дисперсионного анализа книга [264] дает весьма интересныйе и полный обзор по основным методам. Для знакомых с дисперсионным анализом и другими параметрическими методами книга [265] дает довольно детальное сравнение *параметрических* и *непараметрических* методов.

***Непараметрические критерии* (*тесты*).** Как уже отмечалось выше, корректное применение *параметрических* критериев базируется на гипотезе о *нормальном распределении* совокупностей, из которых извлекаются сравниваемые выборки. Однако, это не всегда имеет место. Важное обстоятельство в ряде случаев состоит в необходимости работать не только с *количественными* переменными, но и с *качественными*, многие из которых выражаются последовательностями чисел, индексами и другими условными знаками. В таких случаях необходимо использовать *непараметрические* критерии, которые называются также ***критериями без распределения***.

На сегодня известно достаточно много *непараметрических критериев*; среди них существенное место занимает так называемый ***ранговый критерий,*** применение которого основывается на ранжировании элементов сравниваемых выборок. При этом, сравниваются не сами элементы ранжированных рядов, а их последовательные номера, или *ранги*. Проблемы использования непараметрических критериев, базирующихся на ранговом принципе, достаточно подробно рассматриваются в книге [266]. Тогда как непараметрические критерии в целом достаточно подробно рассматриваются, например, в книгах [133, 258, 259, 265]. В частности, в книге [133] автор попытался представить некоторые из самых простых и полезных непараметрических статистических методов в форме, которая приемлема и как вводный учебник для студентов без предварительных знаний по общей статистике, и как руководство для исследователей с небольшой практикой в статистике.

Необходимо отметить, что непараметрические методы являются наиболее подходящими для студентов в социальных науках, потому что их простота дает достаточно полное понимание стандартных статистических идей, не требуя серьезной математической подготовки. Более того, будучи простыми для понимания, *непараметрические* методы полезны, прежде всего, в приложениях социальных наук, где более строгие предположения о *параметрических* методах часто являются невыполнимыми. Ниже рассмвтриваются два непараметрических критерия, используемые для проверки **Ho**-гипотезы при сравнении как независимых, так и зависимых выборок.

***X-критерий Ван дер Ваардена.*** Данный критерий относится к группе ранговых критериев и используется для проверки **Ho**-гипотезы при сравненим друг с другом *независимых* выборок. Техника вычисления на основе ***X***-критерия состоит в следующем. Сравниваемые выборки *ранжируются* в один общий ряд в порядке увеличения значений переменной. Затем каждому члену этого ряда приписывается серийный **R**-номер, отмечающий его положение в общем *ранжированном* ряде. Далее, на основе серийных **R**-номеров одной из выборок, как правило меньшей по размеру, вычисляются отношения **R/(N+1),** где **N = n1 + n2** (**n1** и **n2** — *размеры сравниваемых выборок*).

С помощью специальной таблицы (*например*, *табл*. **IX** [155]) вычисляются значения функции **ψ[R/(N+1)]** для каждого значения **R/(N+1)** выборки, наименьшей из сравниваемых выборок. Суммируя полученные результаты (*обязательно с учетом знаков*), мы получаем значение $\mathbf{X_f} = \sum\boldsymbol{\psi}\mathbf{[R/(N+1)]}$**.** Данное **Xf**-значение сравнивается с критической точкой **Xst** ***X***-критерия для принятого уровня значимости (**α**) и общего количества **N** членов сравниваемых выборок, т.е. **N = n1 + n2.** Критические точки ***X***-критерия для уровней значимости (**α = 5%** и **α = 1%)** и для *общего* количества **N** (*с учетом разности* **n1 - n2**) членов двух выборок определяются на основе специальной таблицы (*например*, *табл*. **X** [155]).

При сделанных выше предположениях **Ho**-гипотеза состоит в том, что *сравниваемые* выборки извлечены из генеральных совокупностей с идентичными функциями распределения. Если, однако, окажется, что $\mathbf{X_f \geq X_{st}}$**,** то **Ho**-гипотеза отвергается для принятого уровня значимости.

Для *упрощения вычислений* на основе данного алгоритма в среде *Maple* была создана *процедура* ***X_test_VW,*** возвращающаят 2-элементную последовательность вида {**N, [R/(N + 1)]**}, где **N** — общее количество членов двух сравниваемых выборок и and **[R/(N + 1)]** — список значений отношения **R/(N+1).** В разделе **10.5.3** представлены *описание* и *пример* применения процедуры для анализа двух конкретных выборок.

***U-критерий Манна-Уитни.*** Данный критерий дает хорошую возможность *проверки* гипотезы о *принадлежности* сравниваемых *независимых* выборок к той же самой *генеральной совокупности* или к совокупностям с идентичными параметрами (*средние* и *дисперсии*). Критерий относится к *ранговым*, как и предыдущий ***X****-критерий*. Алгоритм вычислений согласно критерию *Манна-Уитни* состоит из следующих трех этапов, а именно.

На ***первом*** этапе числовые значения сравниваемых выборок помещаются в возрастающем порядке в один общий ряд с единой индексацией от **1** до **N = n1 + n2,** где **n1** и **n2** — размеры выборок. Номера значений в общем ряде и определяют их *ранги*.

На ***втором*** этапе для каждого из этих выборок вычисляются суммы **R1** и **R2** рангов, а затем вычисляются две величины **U1** и **U2** согласно следующим формулам:

$$\mathbf{U1 = R1 - \frac{n1*(n1+1)}{2}}, \qquad \mathbf{U2 = R2 - \frac{n2*(n2+1)}{2}} \tag{54.6}$$

Данные две формулы отражают отношение между суммами рангов *первой* и *второй* выборок. Наконец, на ***третьем*** этапе в качестве **U**-критерия выбирается величина $\mathbf{U_f = \mathit{min}\ \{U1, U2\}}$, котораяй сравнивается с табличным критическим значением **Ust.** Критические точки **Ust** для **n1, n2** и принятого уровня значимости (**α**) определяются из специальной таблицы (*например*, *табл*. **XI** [155]). Если имеет место соотношение $\mathbf{U_f > U_{st}}$, то мы можем принять **Ho**-гипотезу для заданного уровня значимости (**α**); в противном случае гипотеза отвергается.

В целях упрощения оценки достоверности **Ho**-гипотезы посредством непараметрического **U**-критерия Манна-Уитни в среде *Maple* была реализована процедура ***U_test_MW***, которая возвращает 3-элементный список вида **[n1, n2, Uf]**, а именно: **n1** — количество членов *первой* выборки, **n2** — количество членов *второй* выборки, и **Uf** — *фактическое* значение **U**-*критерия*. В разделе **10.5.3** представлены описание и пример применения процедуры для анализа двух выборок, идентичных рассмотренным в случае предыдущего критерия.

Для получения дальнейший опыт в использовании *непараметрических* критериев читателю рекомендуется книга [133]. Данная книга содержит ряд наиболее простых и самых полезных *непараметрических* статистических методов в форме, которая приемлема и как вводный курс для студентов без *предварительных* знаний по статистике, и как *справочник* для исследователей с небольшой практикой в статистике. Если же вы имеете желание *более глубоко* ознакомиться с *непараметрическими* методами, предложенными в настоящей книге, то рекомендуются книги [265, 266], с которым вы должны быть способны теперь справиться. Для хорошего введения в некоторые из более сложных способов анализа категориальных данные может быть довольно полезна книга [267]. Наконец, налучший способ познать статистические критерии состоит в том, чтобы использовать их в практическом контексте, например, на основе *Maple*-процедур, представленных в настоящем разделе для статистической проверки гипотез, а также в главе **10** наряду с нашими книгами [144, 302, 303].

## 7.6. Элементы корреляционного и регрессионного анализа

Так как многие явления находятся во взаимной связи, то изучение их важнейшая задача статистики, решаемая ею согласно собственной методологии, определяемой характером исходных данных и целей исследования. Для получения адекватных реалиям результатов статистика требует исследования конкретного явления не изолированно, а во взаимосвязи с основными влияющими на него факторами и другими явлениями. Взаимосвязь явлений – весьма многоаспектное понятие, различающееся ее видами и формами. По характеру зависимости различаются *функциональные* (*полные*) и *корреляционные* (*неполные*) связи. При этом, основным видом связи, изучаемым статистикой, является ***корреляционная***, проявляемая только в массовых случаях (*в среднем*). В отличие от корреляционной, функциональная связь определяет значение результативного признака по значению факторного для каждого конкретного случая. Известно, что установление функциональной связи, в общем случае – значительно более сложная задача, чем корреляционной, которая в ряде случаев может быть определена с той или иной степенью точности даже путем логического анализа и просто здравого смысла.

По *направленности* связи разделяются на *прямые* (*положительные*) и *обратные* (*отрицательные*), а по их *аналитическому* выражению – на *линейные* и *нелинейные*. ***Прямая*** связь характеризуется тем, что с ростом значения *факторного признака* растет (*в среднем*) и значение *результативного*, для ***обратной*** связи – наоборот. Различают также связи ***непосредственные*** и ***косвенные:*** фактор **X** может оказывать влияние на **Y**-фактор *непосредственно* [*формально:* **Y = Y(X)**] или *косвенно* через некоторый промежуточный **Z**-фактор [*формально:* **Y = Y(Z(X))**]. Связи между явлениями могут быть ***слабыми*** и ***сильными*** (*тесными*); средства корреляционного анализа

дают возможность оценивать и данный показатель связи явлений. Связь двух признаков называется ***парной*** (*или просто*) ***корреляцией,*** тогда как влияние нескольких факторов на результативный – ***многофакторной*** (*множественной*) ***корреляцией***. В данном разделе нами будут рассмотрены базовые элементы анализа, в основном, парных корреляции и регрессии. Известно, что статистические показатели обусловлены многими факторами и причинами, связаны определенными зависимостями, исследование которых имеет большое научное и прикладное значение. На основе теоретического анализа статистика устанавливает наличие и направленность взаимосвязей явлений, измеряет и выражает их посредством специальных *показателей связи* и *функциональных уравнений*. Элементы данного анализа рассматриваются несколько ниже.

Для исследования взаимосвязи явлений статистика использует ряд приемов и методов. Достаточно сложные типы взаимосвязи могут быть установлены на основе статистических группировок данных, о которых говорилось в главе **4.** *Группировки* дают возможность выявлять наличие или отсутствие зависимости *результативного* признака от *факторных*, положенных в основу группировки. Однако, с помощью группировочного метода можно характеризовать только самые общие черты связи, но получать более точные показатели такой связи этот метод, вообще говоря, не дает возможности. Вместе с тем, *методу группировок* принадлежит большая роль в исследовании связей между *количественными* и *качественными* признаками.

Метод ***аналитических группировок*** определяет влияние *качественного* признака на значения: *относительных*, *средних величин*, *показателей вариации количественных* признаков. Основной *принцип* изучения связи *группировочным методом* состоит в том, что в качестве *группировочного* признака выбирается *факторный*, а в качестве *сказуемого* статистической таблицы выбираются *средние* значения (*относительные* или *абсолютные*) одного или нескольких *результативных* признаков. Тогда изменение факторного признака при переходе от одной группы к другой группе вызывает соответствующее изменение (*если связь имеется*) результативного признака. При искушенности исследователя и владении им в достаточной мере как самим методом, так и сущностью изучаемого явления, данный метод позволяет решать достаточно сложные задачи на установление связи между явлениями. Например, группировка публикаций **ТТГ** по их типу и месту издания (табл. **1**) позволяет получать ряд весьма интересных *зависимостей* между *средними* величинами некоторых их показателей и *факторными* признаками. Следует иметь в виду,что неверный выбор признаков может приводить к неправильным результатам анализа связи между признаками.

В отличие от большинства других *исследовательских* сфер деятельности, в которых *взаимосвязь* явлений устанавливается экспериментально или теоретически, статистика для этих целей использует, как правило, полученные тем или иным способом статистические данные. В этом случае статистические методы позволяют определять характер влияния *факторного* признака на *результативный*. Наиболее известными из данных методов, наряду с группировочными, являются: *корреляционный*, *регрессионный*, *факторный* и *дисперсионный анализы*. Мы остановимся на двух первых, решающих следующие основные задачи:

1. *определение аналитической связи между вариациями признаков* **X** *и* **Y** (*уравнение*, ***модель регрессии***)
2. *вычисление степени связи между признаками* (***коэффициент корреляции*** *и* ***корреляционное отношение***).

*Регрессионный* и *корреляционный* методы позволяют *количественно* и *аналитически* исследовать влияние факторов на изучаемое явление, однако применение этих методов требует четких целенаправленности и качественного анализа полученных результатов. В противном случае

не исключены ситуации, когда методы дают отрицание наличия (*и/или слабую степень*) связи между признаками при их *сильной взаимосвязи* на самом деле, и *наоборот*. Для эффективности *обоих* методов необходимо, чтобы исследователь хорошо владел не только ими как таковыми, но и самим объектом исследования в целом.

Понятие *регрессионной зависимости* является частным случаем более общего понятия – ***стохастической зависимости:*** *переменная* **Y** *находится в стохастической зависимости от* **X,** *если каждому значению* **X** *соответствует ряд распределения* **Y** *и с изменением* **X** *эти ряды закономерно изменяются*. Если же ряды не изменяются либо изменяются случайным образом, то **Y** не зависит от **X.** Основная задача *регрессионного анализа* состоит в обнаружении факторов, влияющих на исследуемое явление, и построения его регрессионной модели связи. Метод регрессионного анализа состоит из ряда этапов, а именно:

1. *постановка задачи и выбор результативных и факторных признаков исследуемого явления*
2. *сбор статистических данных для анализа и их верификация с целью проверки достоверности*
3. *предварительный анализ связи* (*группировочный*, *графический*, *компьютерный методы и ряд др.*)
4. *исследование парных и многофакторных связей между явлениями*
5. *оценка достоверности результатов анализа и их интерпретация*.

В экономико-статистической практике весьма часто постулируют простую ***модель линейной регрессии*** (**ЛМР**), когда функциональная связь между явлениями **X** и **Y** имеет вид **Y(X) = A*X + B,** где **A, B** – некоторые постоянные величины (*параметры*), вычисляемые по *методу наименьших квадратов* (**МНК**) либо иным методом. *Уравнение регрессии* характеризует изменение среднего уровня **Y**-признака в зависимости от факторного **X**-признака. Оно определяет математическое ожидание групповых средних **Y**-признака при различных значениях **X**-признака. В **ЛМР** результативный **Y**-признак изменяется равномерно под влиянием факторного **X**-признака; модель имеет весьма широкое применение, ее параметры **A** и **B** легко вычисляются и интерпретируются, однако в реалиях такой тип связи является относительно редким. Поэтому **ЛМР** часто рассматривают как некоторое упрощение реальной связи. В ряде случаев путем логарифмирования и некоторые *нелинейные* модели регрессии сводимы к *линейной*, что позволяет упрощать вычисления их параметров посредством, например, **МНК.**

В общем случае *линейная модель регрессии* (**ЛМР**) описывается уравнением следующего вида:

$$Y(X) = A_0*F_0(X) + A_1*F_1(X) + A_2*F_2(X) + \dots A_n*F_n(X) \qquad (55)$$

где $A_k$ – параметры, определяемые посредством, как правило, **МНК;** $F_k(X)$ – известные функции от **X**-переменной и **n** – *порядок* модели (**k = 1 .. n**). Параметры $A_k$ **ЛМР** вычисляются посредством **МНК** по результатам ряда наблюдений ($X_1, X_2, \dots, X_m$) и ($Y_1, Y_2, \dots, Y_m$) с очевидным условием (**m > n**). К классу **ЛМР** относятся и более общие *модели* регрессии, чьи функциональные уравнения имеют следующий относительно простой вид, а именно:

$$G(Y) = A_0*F_0(X) + A_1*F_1(X) + A_2*F_2(X) + \dots A_n*F_n(X) \qquad (56)$$

в случае, когда **G(Y)** имеет *обратную* функцию. Наиболее применяемые модели *регрессионного* статистического анализа имеют следующий вид:

**1)** $Y(X) = A_1*X+A_0 \qquad Y(X) = \sum_{k=0}^{n} A_k*X^k \qquad Y(X) = A_1*\ln(X)+A_0$

$$Y(X) = \frac{A_2}{X} + A_1*X+A_0 \qquad \ln(Y) = \sum_{k=0}^{n} A_k*X^k \qquad \frac{1}{Y} = \sum_{k=0}^{n} A_k*X^k$$

**2) $F_k(X)$** *в формулах* (**55, 56**) **–** *ортогональные полиномы* (*например*, *полиномы Чебышева*)

**3) $F_k(X)$** *в формулах* (**55, 56**) *принадлежат классу сплайн-функций*

Для получения значений неизвестных $A_k$-*параметров* **ЛМР** используется, как правило, **МНК,** который в отличие от других подобных методов (*методы моментов*, *максимума правдоподобия и др*.) не требует *априорной* информации о виде распределения, что очень важно на начальных этапах регрессионного статистического анализа. Данный метод весьма широко используется в математике и в ее многочисленных приложениях. Ознакомиться с его приложениями для различного типа сглаживания данных можно, например, в книгах [59, 66, 74, 94, 150, 156, 161, 186, 207, 214, 243-246, 278]. Так, последние четыре книги достаточно подробно рассматривают теоретические и прикладные вопросы регрессионного анализа регресса. Тогда как детальное обоснование статистических моделей можно найти в книгах [252-255]. Здесь мы ограничимся простейшим случаем **ЛМР,** описываемой линейным уравнением следующего вида:

$$Y(X) = A*X + B \tag{57}$$

В применении к **ЛМР** (57) использование **МНК** для определения двух ее параметров **A** и **B** по наблюдениям $(X_1,Y_1)$, $(X_2,Y_2)$, ..., $(X_n,Y_n)$ состоит в *минимизации функции* **Z(A, B)** с *последующим* вычислением *параметров* из полученной *системы двух линейных относительно* **A** и **B** *уравнений*:

$$Y_k = A+B*X_k \qquad Z(A,B) = \sum_{k=1}^{n} (Y_k - Y(X_k))^2$$

$$\begin{cases} \dfrac{\partial Z(A,B)}{\partial A} = 0 \\ \dfrac{\partial Z(A,B)}{\partial B} = 0 \end{cases} \qquad \begin{cases} \sum\limits_{k=1}^{n} X_k*(Y_k - A*X_k - B) = 0 \\ \sum\limits_{k=1}^{n} (Y_k - A*X_k - B) = 0 \end{cases}$$

$$\begin{cases} M[X^2]*A + M[X]*B = M[X*Y] \\ M[X]*A + B = M[Y] \end{cases} \tag{58}$$

$$A = \frac{M[X*Y] - M[X]*M[Y]}{D[X]} \qquad B = \frac{M[Y]*M[X^2] - M[X]*M[X*Y]}{D[X]}$$

где **M[X]** и **D[X]** – соответственно *математическое ожидание* и *дисперсия* случайной **X**-*величины*; в случае наблюдений $\{(X_k, Y_k)\}$ ($k = 1 .. n$) параметр **M[X]** обычно полагается равным средней арифметической **X**-величины. При этом, величины **M[Y], M[XY], $M[X^2]$** и $D[X]=M[X^2] - M^2[X]$ вычисляются подобным образом (*см. раздел* **2.3**). Предоставляем читателю в качестве весьма полезного упражнения детально провести вычисления (58) для получения значений параметров **A** и **B,** подставив которые в (57), получаем конкретную реализацию **ЛМР** для исследуемого явления по полученным статистическим данным $\{(X_k, Y_k)\}$ ($k = 1 .. n$) его наблюдения. Пример разработки **ЛМР** в *Maple*-среде представлен на рис. **12** с тем отличием, что для вычисления параметров **A** и **B** модели используется статистическая процедура ***fit*** пакета, чей алгоритм реализован на основе **ЛМР**. На данном рисунке представлены прямая линия регрессии и точки, соответствующие статистическим данным наблюдения. В качестве наблюдения были отобраны *ежегодные* объемы публикаций **ТТГ** (*монографического характера и в целом*) за период 1990 – 1999 (*табл*. **7,** *графы* **3** и **9** *соответственно*).

```
> Xval:= [762, 702, 561, 1156, 115, 450, 558, 1192, 1196, 3275]: Xval:= sort(Xval): Yval:= [785, 709, 561,
  1170, 115, 535, 629, 1251, 1266, 3354]:  Yval:= sort(Yval):  with(stats):  n:= nops(Xval): F:= [TIMES,
  BOLD, 9]:  LRM:= fit[leastsquare[[X, Y], Y = A*X + B], {A, B}]([Xval, Yval]);
```

LRM := Y = 70172055*X/68943901 - 1588810069/68943901

```
> A:= diff(rhs(%), X),  B = subs(X = 0, rhs(%));
```

A:= 70172055*X/68943901,  B = -1588810069/68943901

```
> `M[Xval]` = describe[mean](Xval),  `M[Yval]` = describe[mean](Yval);
```

M[Xval] = 9967/10,  M[Yval] = 2075/2

```
> P1:=plot(%%[1]*X+rhs(%%[2]), X=Xval[1]..Xval[n], thickness=2, color=black): L:=[seq([Xval[k],
  Yval[k]], k=1..n)]:  with(plots):  P2:=pointplot(L, color=black, thickness=3, symbol = CIRCLE):
  display([P1, P2], axesfont = F, labels = [`(X, Xval)`, `(Y, Yval)`], labelfont = F);
```

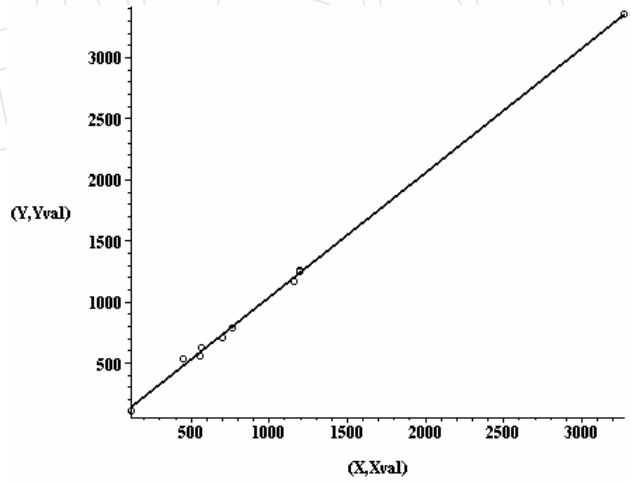



**Рис. 12. Построение в *Maple*-среде ЛМР, базирующейся на данных табл. 7**

Данный рис. показывает, что линия регрессии очень хорошо согласуется с точками данных изучаемого наблюдения. Для понимания примера (рис. **12**) читатель может ограничиться информацией по пакету *Maple*, например, в объеме книг [97, 98, 139-141, 143, 144]. В качестве полезного упражнения читателю рекомендуется вычислить на основе **МНК** параметры **A, B** и **C** для следующих *нелинейных моделей регрессии* (**НМР**):

$$\mathbf{y(x) = a*x^2 + b*x + c} \quad \mathbf{y(x) = \frac{a}{x} + b*x + c} \quad \mathbf{ln(y) = a*ln(x) + b}$$

Логико-экономическая интерпретация параметров модели регрессии проводится на основе анализа сущности исследуемого явления. Так, например, **A**-параметр модели говорит о том, на сколько в *среднем* для всех наблюдений изменится значение **Y** при изменении **X**-фактора на *одну единицу* в пределах установленной наблюдением вариации признака.

Для экономической интерпретации *линейных* и *нелинейных* моделей *парной* регрессии удобно пользоваться так называемым ***коэффициентом эластичности*** (**КЭ**), который в случае **ЛМР** вычисляется по простой формуле **КЭ = A*M[X]/M[Y]**. Данный коэффициент показывает, на сколько процентов в среднем изменяется величина **Y** с изменением значения **X**-фактора на **1%.** Так, для случая примера (рис. **12**) *коэффициент эластичности* равен величине **КЭ = A*** **M[Xval]/M[Yval]** = 0.978, т.е. величина **Yval** изменяется на 0.978% при изменении величины **Xval** только на **1%.** Таким образом, значение **КЭ** говорит о наличии однородной линейной зависимости между *переменными* **Xval** и **Yval,** т.е. **Xval ≈ Yval + *const*.** Таким образом, *уравнение регрессии* позволяет оценивать *среднее* значение *результативного* признака и аналитически отражать вариацию ожидаемых *групповых средних*, для чего вполне достаточно в него вместо признаков **X** и **Y** подставлять их *средние* $\overline{\mathbf{X}}$ и $\overline{\mathbf{Y}}$ значения.

При наличии взаимосвязи между *факторными* признаками результаты *регрессионного* анализа могут инициировать ошибочные выводы, поэтому выявление такой связи и устранение ее

влияния на окончательные результаты является важной задачей регрессионного анализа. В следующей главе подобная задача сводится к устранению влияния факторов *автокорреляции* и *авторегрессии* при изучении взаимосвязи динамических рядов. Наконец, следует отметить, что теория *нелинейной регрессии* разработана относительно слабо и модели для нее во многих случаях строятся эмпирически, т.к. *качественный* анализ далеко не всегда позволяет выбирать удовлетворительную форму кривой регрессии. Здесь весьма эффективным средством может оказаться описываемый в книге *компьютерный* подход к выбору формы кривой по известным точкам статистических данных наблюдения.

Главнейшей задачей ***корреляционного*** *анализа* является установление *тесноты связи* между *результативным* и *факторным* признаками, т.е. степени влияния **X**-признака на **Y**-признак. Для этой цели используется целый ряд показателей, из которых для случая **ЛМР** наиболее известным является *коэффициент корреляции* **CC(X, Y)** *между* **X-** и **Y-***признаками*, вычисляемый по следующей общей формуле:

$$\mathbf{CC(X,Y)} = \frac{\mathbf{M[X*Y]} - \mathbf{M[X]*M[Y]}}{\sqrt{\mathbf{D[X]*D[Y]}}} \tag{59}$$

где **M[X]** и **D[X]** – соответственно *математическое ожидание* и *дисперсия* случайной **X**-*величины*. Сопоставляя формулы (59) и (58) для **A**-параметра **ЛМР,** можно легко получить следующее полезное соотношение между ними:

$$\mathbf{CC(X,Y)} = \mathbf{A}*\sqrt{\frac{\mathbf{D[X]}}{\mathbf{D[Y]}}} \tag{60}$$

В частности, из формулы (60) и соотношения (16.3) для дисперсии следует, что **CC(X, Y) = 1,** если статистические данные **{($X_k$, $Y_k$)}** (**k = 1 .. n**) строго соответствуют **ЛМР** линейного вида **Y(X) = A*X + B.** В общем же случае имеет место соотношение **| CC(X, Y) | ≤ 1**. Показатель **CC(X, Y)** определяет силу и направленность связи между **X-** и **Y-**признаками для случая **ЛМР,** а именно**:** (**1**) *при* **CC(X, Y) > 0** [**CC(X, Y) < 0**] *с ростом значения* **X** *растет* [*убывает*] *значение* **Y;** (**2**) *при попадании значения* **CC(X, Y)** *в интервал* [**0, 0.1**], [**0.1, 0.3**], [**0.3, 0.7**] *или* [**0.7, 1.0**] *примерную степень связи между* **X-** *и* **Y-***признаками можно соответственно определять как нулевую, слабую, среднюю или сильную.*

***Среднеквадратическая ошибка*** (**ErCC**) выборочного **CC**-показателя парной корреляции вычисляется по весьма простой формуле $\mathbf{ErCC} = (1 - \mathbf{CC}^2)/\sqrt{n-1}$, которая непосредственно используется для вычисления доверительного интервала для **CC**-показателя. Согласно современным данным свесьма большой уверенностью можно утверждать, что значение **CC**-*показателя* при достаточно *большом* объеме **n** статистических наблюдений должно превышать **ErCC**-величину не менее, чем в **3** раза, т.е. **CC ≥ 3*ErCC**. При *невыполнении* этого неравенства наличие связи между исследуемыми явлениями нельзя считать доказанным.

Для случая *нелинейной модели регрессии* (**НМР**) линейный показатель **CC(X, Y)** теряет свой смысл и для измерения степени связи между **X-** и **Y-**признаками весьма часто используется *корреляционное отношение* (**CR**), вычисляемое по следующей общей формуле:

$$\mathbf{CR(X,Y)} = \sqrt{\frac{\mathbf{D[Y]} - \mathbf{D[Y(X)]}}{\mathbf{D[Y]}}} \tag{61}$$

где **D[Y]** – *дисперсия* **Y**-данных наблюдения и **D[Y(X)]** – *дисперсия*, вычисляемая по **X**-данным наблюдения на основе выбранной **Y(X)-**модели регрессии, т.е. вычисляется величина:

$$D[Y(X)] = \frac{\sum_{k=1}^{n} (Y_k - Y(X_k))^2}{n}$$

Значение показателя **CR(X, Y)** лежит в интервале [**0, 1**] (*с учетом алгебраического знака: от* **-1** *до* **+1;** *знак указывает лишь на направленность связи*); большее его значение говорит о большей степени корреляционной зависимости.

При **CR(X, Y) = 0** между признаками **X** и **Y** не существует корреляции, а при **CR(X, Y) = 1** между ними имеется функциональная связь. Показатель **CR(X, Y)** является мерой связи как для *линейной*, так и для *нелинейной* формы связи между явлениями. При этом, для *нелинейной* формы **CR**-показатель является на сегодня *единственно* общепринятым измерителем степени связи; **CR**-показатель также иногда называют ***индексом корреляции.*** При этом, **б**олее детально обсуждение **CC-** и **CR**-показателей будет проведено в конце раздела настоящей главы.

Для проверки гипотезы на соответствие статистических данных **{($X_k$, $Y_k$)}** (**k=1..n**) выбранной **ЛМР** можно использовать следующий простой критерий: *в случае справедливости неравенства* **n*(CR$^2$(X, Y) - CC$^2$(X, Y)) < 11.37,** *где* **n** *– число точек данных, можно считать вполне приемлемой* **ЛМР.** Рассмотрим следующий несложный пример (рис. **13**) разработки *регрессионной модели* и вычисления показателей **CC** и **CR.**

```
> A:= [0,33,38,42,76,94,96,98,102,125,145,174,186,205,220,222,236,240,266,284,292,304,314,316,320,
  336, 345,384, 395,405]:  U:= [0,0,0,5,7,12,13,14,14,15,18,27,24,22,29,42,48,56,64,68,68,70,65,58,55,60,
  63,67,75,95]:  Digits:= 4:  LRM_NRM(U, A, NRM, F, CR),  CR,  SimpleStat:- CC(U, A);
                    -5.999 + 0.00008486*X^2 + 0.1886*X,  0.9599,  0.9593
> F;
```

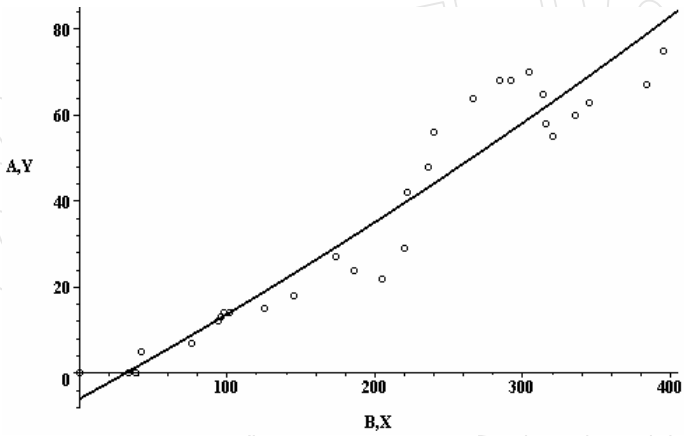



**Рис. 13. Разработка *нелинейной модели регрессии* с вычислением *коэффициента корреляции* (CC) и *корреляционного отношения* (CR)**

В примере (рис. **13**) исследовалась зависимость между цитируемостями научных работ **ТТГ** по **МТОС** в СССР (**U**) и *за рубежом* (**A**); при этом, был исключен фактор *автокорреляция* между уровнями динамических рядов **U** и **A.** Вопросы автокорреляции временных (*динамических*) рядов рассматриваются в следующей главе настоящей книги.

Для разработки **НМР** для случая наших начальных данных – *результирующей* **U-**переменной и *факторной* **A**-переменной, соответствующих данным граф **4** и **5** из табл. **9** соответственно – используется пакет *Maple*. В *первой* части *Maple-программы* на основе *выбранных* статистических

данных формируются списки **U** и **A,** соответствующие вышеупомянутым графам (рис. **13**). Для разработки **НМР** в соответствии с заданными *результирующей* и *факторной переменными*, а также для вычисления *корреляционного отношения* (**CR**) с *общим графическим* представлением входных данных и разыскиваемой модели регрессии используется специальная процедура ***LRM_NRM***. Описание процедуры может быть найдено в разделе **10.5.2** настоящей книги.

Процедура непосредственно возвращает требуемую *линейную* или *нелинейную* (*квадратичную*) однофакторную модель регрессии в виде *уравнения корреляции*, тогда как через два последних аргумента возвращает график *квадратичной* модели регрессии вместе с исходными данными, а также *корреляционное отношение*. Фрагмент (рис. **13**) представляет пример применения ***LRM_NRM*** процедуры для разработки нелинейной модели регрессии для вышеупомянутой результирующей **U**-переменной и факторной **A**-переменная. Наряду с этим, данный пример вычисляет коэффициент корреляции между этими переменными на основе **CC**-процедуры, описанной в разделе **10.5.4** настоящей книги.

В результате выполнения *Maple*-документа возвращается искомая *однофакторная* нелинейная модель регрессии, а именно: **0.00008486*X^2 + 0.1886*X - 5.999**. Наряду с этим, вычисляются значения **CC(U, A) =** 0.9593 и **CR(U, A) =** 0.9599 для *коэффициента корреляции* и *корреляционного отношения* соответственно. В завершении фрагмента (рис. **13**) через **F**-аргумент процедуры ***LRM_NRM*** возвращается общий график искомой модели регрессии и распределения точек исходных данных (**U, A**). В свете вышесказанного о *регрессии* и *корреляции* изучение влияния факторной **A**-переменной на результирующую **U**-переменную несомненно подтверждает, что таковая в значительной степени соответствует действительности. Полученные значения *коэффициента корреляции* и *корреляционного отношения* говорят о существовании достаточно тесной связи между вышеупомянутыми переменными **U** и **A.**

Разработка моделей регрессии базируется на вышеупомянутом *методе наименьших квадратов*, тогда как выше оба показателя *корреляции* были рассчитаны на основе классических формул. Для понимания *Maple*-фрагмента читатель может ограничиться информацией по пакету, например, в объеме книг [97, 98, 139-141, 143, 144]. Более детально статистические средства *Maple* рассматриваются в последней главе настоящей книги и в нашей книге [302, 303].

Показатель **CR(X, Y)** является более совершенной мерой степени стохастической связи между случайными переменными **X** и **Y,** чем **CC(X, Y)**. Коэффициент корреляции является вполне надежной мерой связи в том случае, когда он численно равен или весьма близок к значению **CR(X, Y)**. В противном случае мы в той или иной мере недооцениваем степень связи, измеряя ее посредством показателя **CC(X, Y)**. Параметры модели регрессии, показатели **CC(X, Y)** и **CR(X, Y)**, как правило, определяются по выборочным данным, что отличает их значения от соответствующих значений *генеральной* совокупности. Поэтому необходимо уметь определять их точность и границы для их доверительных интервалов. Подробнее с этими вопросами можно ознакомиться, например, в книгах [92, 112, 123-125, 132, 155, 157, 268, 269, 275, 284, 285, 301, 305-311].

Наряду с *коэффициентом корреляции* (**CC**) в целом ряде случаев используется так называемый *коэффициент детерминации* (**DC**), вычисляемый как **DC = CC^2**. Показатель **DC** характеризует долю дисперсии результативного признака, объясняемую влиянием аргументов, входящих в **ЛМР**. В этом случае **DC**-показатель совпадает с квадратом **CR**-показателя.

Наконец, под ***множественной*** *регрессией* понимается *исследование* статистической зависимости *результативного признака* от *нескольких* (*более одного*) *факторных*. При этом, методы *исследования* функциональной зависимости здесь, в общем случае, непригодны, когда вопрос касается стохастической связи. Основная задача *многофакторного анализа* сводится к вычислению значения результативного **Z**-признака по значениям факторных $\mathbf{X_k}$-признаков

(**k = 1 .. n; n > 1**). Построение *многофакторной модели регрессии* (**МФМР**) включает в себя: (**1**) *выбор основных факторных признаков;* (**2**) *выбор типа модели* и (**3**) *определение числа наблюдений, необходимого для получения несмещенных оценок*. Выбор основных *факторных* признаков должен отвечать ряду требований, а именно:

- *они должны оказывать наиболее существеное влияние на результативный признак*
- *количество факторных признаков следует выбирать* (*без искажения картины исследуемого явления*), *по-возможности*, *минимальным*
- *факторные признаки не должны находиться между собой в тесной или функциональной связи, в противном случае результаты анализа могут быть недостоверными*
- *отбор некоррелируемых между собой признаков производится путем вычисления между ними показателей корреляции* **CC** *и/или* **CR.**

Тип модели выбирается как исходя из сущности исследуемого явления, так и принимая во внимание максимальную простоту модели. На практике наибольшее распространение получили *линейные* и *приведенные к линейным* формам связи модели по целому ряду причин практического и содержательного характера. При этом, особой внимательности требует отбор статистических данных для проведения *многофакторного* анализа. При проведении анализа часто приходится иметь дело с малыми выборками; если число *факторных признаков* достаточно велико, то это приводит к существенному увеличению размеров доверительных интервалов для параметров **МФМР** и уменьшению их достоверности. Статистическая практика исходит из следующего эмпирического правила: *при построении* **МФМР** *число статистических данных наблюдения в выборке должно превышать число факторных признаков в* **8** *раз*. Не нарушая общности и ради простоты, кратко рассмотрим случай линейной **МФМР** с *двумя факторными* признаками, уравнение которой имеет следующий довольно простой вид:

$$\mathbf{Z(X, Y) = A*X + B*Y + C} \tag{62}$$

Для вычисления параметров **A, B** и **C** этой модели используется, как правило, **МНК,** т.е. они находятся как решение системы линейных уравнений следующего вида:

$$\mathbf{F(A,B,C)} = \sum_{k=1}^{n} (z_k - A*x_k - B*y_k - C)^2; \quad \begin{cases} \dfrac{\partial F(A,B,C)}{\partial A} = 0 \\ \dfrac{\partial F(A,B,C)}{\partial B} = 0 \\ \dfrac{\partial F(A,B,C)}{\partial C} = 0 \end{cases} \tag{63}$$

Решение *линейной* системы (63) относительно параметров **A, B** и **C** этой модели не составляет особого труда, в частности, классическими средствами линейной алгебры. При наличии **ПК** для этих целей можно воспользоваться математическими пакетами такими как *MathCAD* [14, 15, 17, 18, 29, 127, 128], *REDUCE* [22], *Mathematica* [99, 134-137, 156], *Maple* [97, 98, 139-144, 158, 230, 233, 302, 303] и другими [34, 36, 38, 39, 42, 45, 100-102, 165, 229, 235, 304]. Читателю в качестве полезного упражнения предлагается решить в алгебраическом виде систему (63). Мы же в качестве иллюстрации приведем простой пример построения линейной **МФМР** типа (63), для которой признаками **Z, X** и **Y** являются соответственно ежегодная цитируемость работ **ТТГ** по **МТОС** в целом (табл. **9;** графа **7**), число публикаций группы в СССР и за рубежом (табл. **9;** графы **2** и **3**) за период 1970 - 1999 г.г. При этом, решение задачи проводится в среде математического пакета *Maple* (рис. **14**). Существующие на сегодня многочисленные **ППП** позволяют легко решать эту и подобные ей задачи пользователям, не имеющим специальной программистской подготовки. И выбор здесь пакета *Maple* не обусловлен какими-либо статистическими соображениями.

```
> SU:= [3,9,4,2,3,2,4,0,1,0,3,6,3,4,0,4,1,1,3,1,2,3,1,3,1,1,2,3,5,5]:  AB:= [0,0,0,0,3,1,1,0,1,1,0,2,2,1,3,1,3,
  2,8,2,1,1,0,0,0,1,1,2,5,5]:  Q:= [0,33,38,47,83,106,109,112,116,140,163,201,210, 227,249,264,284,296,
  330,352,360,374,379,374,375,396,408,455,470,500]:  Digits:= 5:  with(SimpleStat):
> `Averages for Q, SU and AB `,  map(evalf, [SR(Q), SR(SU), SR(AB)]);
                    Averages for Q, SU and AB,  [248.37, 2.6667, 1.5667]
> `Dispersions for Q, SU and AB `,  map(evalf, [Ds(Q), Ds(SU), Ds(AB)]);
                    Dispersions for Q, SU and AB,  [20665., 3.7556, 3.2456]
> `Correlation coefficients `,  map(evalf, [CC(SU,AB), CC(Q,SU), CC(Q,AB) ]);
                    Correlation coefficients,  [0.11139, -0.11454, 0.41094]
> with(linalg):  n:= nops(SU):  E:= [seq(1, k = 1 .. n)]:
> S:= (a, b) -> sum(a[k]*b[k], k=1 .. nops(a)):  d:= matrix(3, 3, [S(SU, SU), S(SU, AB), S(SU, E),
  S(SU, AB), S(AB, AB), S(AB, E), S(SU, E), S(AB,E), n]):  B:=vector(3, [S(Q,SU),S(Q,AB),S(Q,E)]):
> R:= evalf(linsolve(d, B)):  Z:= (X, Y) -> R[1]*X + R[2]*Y + R[3]:
> `Two-factor regression model `,  `Z(X, Y)` = Z(X, Y);
                    Two-factor regression model,  Z(X, Y) = -12.042 X + 34.234 Y + 226.85
> PCC(a, b, c) = (CC(a, b) - CC(a, c)*CC(b,c))/sqrt(1 - CC(a, c)^2*(1 - CC(b, c)^2));
```

$$\mathbf{PCC}(a, b, c) = \frac{\mathbf{CC}(a, b) - \mathbf{CC}(a, c)\,\mathbf{CC}(b, c)}{\sqrt{1 - \mathbf{CC}(a, c)^2\,(1 - \mathbf{CC}(b, c)^2)}}$$

```
> `Partial correlation coefficients `,  PCC(Q, SU, AB), PCC(Q, AB, SU);
                    Partial correlation coefficients,  -0.17563,  0.42647
> cr:= (a, b) -> evalf(sqrt(1 - Ds(b)/Ds(a))):  `Correlation ratios `,  cr(Q, R[1]*SU + R[2]*AB +
  R[3]*E),  cr(Q, SU + AB);
                    Correlation ratios,  0.89727,  0.99984
> `Multiple correlation coefficient `,  MCC(Q, SU, AB);
                    Multiple correlation coefficient,  0.45986
```

**Рис. 14. Разработка линейной *двухфакторной* модели регрессии в среде *Maple***

Для разработки линейной *двухфакторной* модели регрессии **Z(X,Y)** используется пакет *Maple*. В *первой* части *Maple*-документа на основе статистических данных формируются векторы **Q, SU** и **AB** адекватные вышеупомянутым переменным **Z, X** и (рис. **14**). Затем, *Maple*-документ вычисляет средние (**Sr**), дисперсии (**Ds**) и линейные коэффициенты корреляции (**CC**) для *совокупностей* **Q, SU** и **AB** на основе процедур **SR, Ds** и **CC** программного модуля ***SimpleStat***. Из полученных результатов следует, что между факторными переменными **SU** и **AB,** с одной стороны, и между результирующей переменной **Q** и факторной переменной **SU,** с другой стороны, имеет место практически полное отсутствие какой-либо связи. Тогда как между *результирующей* переменной **Q** и *факторной* переменной **AB** отношение может быть оценено как связь средней степени – $\mathbf{CC(Q, AB)} \approx 0.411$.

На *втором* этапе для оценки параметров **A, B** и **C** линейной **МФМР** на основе весьма простых преобразований системы (63) для наших конкретных условий (*списки значений переменных* **Q, SU** *и* **AB**) получаем данные для вычисления *матрицы коэффициентов* (**d**) и **B**-*вектора свободных членов*. На основе этих данных решается система трех уравнений, линейных относительно искомых переменных **A, B** и **C** на основе процедуры ***linsolve*** пакета *Maple*. В результате этого получаем вектор [**A, B, C**] = [-12.042, 34.234, 226.85], который однозначно определяет искомую линейной *двухфакторной* модели регрессии **Z(X,Y)**, а именно (рис. **14**):

$$\mathbf{Z(X, Y) = -12.042*X + 34.234*Y + 226.85}$$

Читателю рекомендуется более детально рассмотреть данный пример (рис. **14**), который представляет также и самостоятельный практический интерес. При этом, для понимания данного *Maple*-документа читатель может ограничиться информацией по пакету, например, в рамках книг [99, 139, 140-143, 158].

В отличие от *парной модели регрессии* в случае *линейной* **МФМР** используются коэффициенты корреляции *множественный* и *частные*. *Коэффициент частной корреляции* (**PCC**) характеризует влияние факторного признака, входящего в уравнение регрессии, и измеряет степень связи между ним и *результативным* **Z**-признаком при условии, что остальные *факторные* признаки не оказывают на него влияния. Достоинство **PCC** состоит в том, что имеется объективная оценка существенности его отличия от нуля в генеральной совокупности. Формула **PCC** выражается, в частности, через парные **CC** -показатели и для нашего случая **Z(X,Y)-**модели регрессии (62) оба *частных коэффициента* вычисляются по следующим формулам:

$$\mathbf{PCC(Z,X,Y) = \frac{CC(Z,X) - CC(Z,Y)*CC(X,Y)}{\sqrt{(1-CC(Z,Y)^2)*(1-CC(X,Y)^2)}}} \quad \mathbf{(1)}$$

$$\mathbf{PCC(Z,Y,X) = \frac{CC(Z,Y) - CC(Z,X)*CC(X,Y)}{\sqrt{(1-CC(Z,X)^2)*(1-CC(X,Y)^2)}}} \quad \mathbf{(2)} \qquad (64)$$

Так, формула (64.1) определяет **PCC** факторного **X**-признака на результативный **Z**-признак при игнорировании влияния **Y**-признака. При большом количестве факторных признаков используется *матрица коэффициентов корреляции*. *Коэффициент множественной корреляции* (**MCC**) для нашего случая модели регрессии (62) будет вычисляться по следующей формуле:

$$\mathbf{MCC(Z,X,Y) = \sqrt{\frac{A*PCC(Z,X,Y)*\sqrt{D[X]} + B*PCC(Z,Y,X)*\sqrt{D[Y]}}{\sqrt{D[Z]}}}}$$

В частности в примере (рис. **14**), значения для коэффициентов **PCC(Q, SU, AB), PCC(Q, AB, SU)** и **MCC(Q, SU, AB)** равняются 0.176, 0.426 и 0.4599 соответственно. Это говорит о наличии множественной связи положительной средней степени между **Q,** с одной стороны, и **SU, AB,** с другой стороны. Тогда как два **PCC** противоположного знака вероятнее всего обусловлены нелинейным характером связи для каждой факторной переменной, взятой отдельно, что делает результаты их вычисления сомнительными – суммарный результат действия обоих факторов имеет тенденцию, довольно подобную линейной ассоциации. Наконец, результат вычисления *множественного корреляционного отношения* (**CR =** 0.89727) говорит о существовании достаточно тесной связи, что весьма согласовывается с имеющимися данными логического анализа активности **ТТГ,** в значительной степени определяемой ее как монографическими, так периодическими публикациями и изданиями.

Методы количественной оценки результатов регрессионного анализа состоят в подстановке средних значений факторных признаков в уравнение регрессии и в последующей оценке полученного соотношения. Основным обобщающим показателем оценки корреляции является *корреляционное отношение* (**CR**). В случае, если генеральная совокупность, из которой производится *выборка*, не подчиняется *нормальному* или *близкому* к нему закону *распределения*, то используются методы *ранговой корреляции*, разработанные К. Пирсоном и др. Особенно широкое применение *ранговая корреляция* получила в медицинской статистике для *выявления* влияния различных факторов на состояние здоровья человека наряду с довольно широким применением в биометрике и эконометрике [40, 51, 53, 54, 155, 163, 273].

*Корреляционный* и *регрессионный* методы анализа весьма широко применяются в социально-экономических исследованиях, однако они требуют довольно осторожного использования, а полученные на их основе результаты должны получать *удовлетворительную* интерпретацию, подкрепляемую логическим и качественным анализом на основе сущности исследуемого явления. В первую очередь, это относится к *коэффициентам корреляции*, которые улавливают только связи, приближающиеся к простой пропорциональности. При сложных видах связи этот показатель не гарантирует достоверных результатов, отвечающих реалиям. При этом, для уточнения вопроса взаимосвязи явлений рекомендуется использовать несколько методов связного анализа (*параллельных рядов*, *индексный метод*, *многомерные группировки*, *графический* и *компьютерный метод, и др.*). При этом, хорошее знание сущности исследуемых явлений – основа получения достоверных результатов связного анализа статистических данных.

В заключение раздела приведем ряд критических замечаний по оценке степени связи между явлениями средствами корреляционного анализа, рассмотренными выше. Несмотря на реальность установления и оценки связи между явлениями, а также наличия в статистике ряда показателей степени связи, дискуссии на данный предмет не закончились. По данному вопросу существует весьма широкий спектр мнений – от полного отрицания имеющихся мер оценки связности явлений до абсолютизации таких основных показателей связности, как **CC** и **CR**. Еще соображения и результаты А.А. Чупрова, А.А. Кауфмана, С.Н. Бернштейна и др. показали, что **CC**-показатель имеет определенную значимость и дает неплохие результаты при условии наличия связи, приближающейся к *простой пропорциональности*. При сложных видах связи применение его не дает достоверных результатов. Значительно более общим показателем степени связи рассматривается *корреляционное отношение* (**CR**). Наряду с **CC**- и **CR**-показателями связности явлений имеется целый ряд других средств для проведения связного анализа явлений. Рассматривая данный вопрос и ограничившись только наиболее известными **CC**- и **CR**-показателями связности явлений, мы получили некоторые интересные соображения, носящие методологически-прикладной характер.

Принимая во внимание то обстоятельство, что в основе построения обоих показателей (**CC**, **CR**) связности лежит использование *дисперсии* – меры отклонения значений признака от его средней, естественным образом напрашивается следующая простая модель для дальнейших рассмотрений. При определенных условиях *колебательного* (*периодического*) характера явление будет в качестве своей средней иметь некоторую кривую, близкую к прямой линии. В этом случае можно надеяться, что результирующая периодического процесса, приближаясь к прямой линии, будет существенно влиять на степень связности (*между явлением и его* **IMP**), вычисляемой согласно показателям **CC** и **CR**, таким образом, что достоверность получаемых результатов будет достаточно велика, даже при нелинейности самого явления. В этом случае используется чисто формальное свойство периодических процессов. На основе данного соображения нами была исследована *весьма* простая модель, демонстрирующая определение связности (*в действительности нелинейного характера*) явлений *классическими* статистическими средствами посредством **CC**- и **CR**-показателей связности.

В качестве генеральной совокупности определяется множество всех значений периодической функции **Z=M*sin(N*x)** на отрезке [**0, 0.9**], тогда как выборка из нее объемом (**n =** 100) единиц определяется как множество значений **Z**-функции в точках $\mathbf{X_k = p*k,\ p = 1/n}$ и **k = 0 .. (n - 1).** В среде пакета *Maple* (рис. **15, 16**) в предположении об отсутствии *априорной* зависимости между **Z-** и **X-**признаками проводится анализ связи между ними на основе **CC**- и **CR**-показателей.

```
> n, M, N, p, x:= 100, 5, 6, n^(-1), k -> evalf(p*k, 6):   z:= t -> evalf(M*sin(N*x(t)),  6):
> G:= proc(h, p, n) local t;  [seq(h(t)*p(t), t = 1 .. n)]  end proc:  X:= G(x, 1, n):  Z:= G(z, 1, n):
  XX:= G(x, x, n):  ZX:= G(z, x, n):  E:= G(1, 1, n):  F:= [TIMES, BOLD, 9]:
```

```
> map(with, [linalg, SimpleStat, plots]): `CC`:= CC(X, Z);
```

CC := -0.8231509862

```
> M:= matrix(2, 2, [SR(XX), SR(X), SR(X), 1]): B:=vector(2, [SR(ZX), SR(Z)]): r:=linsolve(M, B):
  Y:= r[1]*X + r[2]*E:  u:= v -> r[1]*v + r[2]: u(v);
```

-10.30826023*v + 5.231867746

```
> DY:=sum((Z[k] - Y[k])^2, k=1..n)/n: `CR`:=sqrt((Ds(Z) - DY)/Ds(Z)); Er:=evalf((1 - CC(X,Z)^2)/
  sqrt(n-1)); n*(CR^2 - CC^2), abs(`CC`) - 3*Er;
```

CR := 0.8231509879

Er := 0.03240467586

0.2800*10^(-6), 0.7259369586

```
> P2:= plot(u(v), v=0 .. 0.9, thickness = 2, color = black): P1:= pointplot([seq([X[j], Z[j]], j=1 .. n)],
  color = black, thickness = 3, symbol = CIRCLE): display([P1, P2], axesfont = F, labels =
  [convert("X", symbol), `Z, Y`], labelfont = F);
```

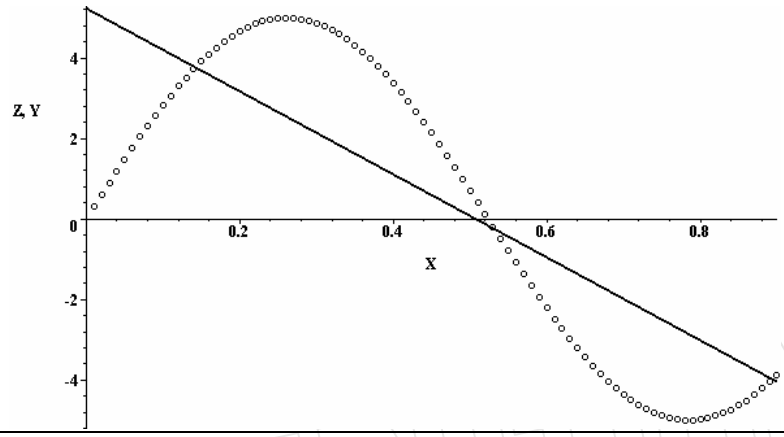



**Рис. 15. Определение связности явлений нелинейного характера посредством *коэффициента корреляции* и *корреляционного отношения*. Проверка гипотезы о *линейном* характере *генерального* распределения на основе критерия согласия Романовского.**

Прежде всего для **CC**-показателя получаем значение **CC** ≈ -0.823151, что говорит о наличии весьма сильной обратной связи в предположении ее линейности. После этого вычисляется для указанных признаков **ЛМР,** и на ее основе вычисляется **CR**-*показатель*, который (*с учетом знака*) дает значение, равное **CC**-*показателю* (*с точностью до* 10^(-6), *что является уже вполне приемлемой степенью идентичности значений с учетом целей нашего исследования*).

Наконец, показывается (рис. **15**), что величина **CC**-показателя *существенно* больше *утроенного* значения (**Er**) средней квадратической ошибки выборочного **CC**-показателя, а именно: **|CC| - 3*Er** ≈ 0.725936. В заключении примера представлены графики значений самой выборки и соответствующей ей **LRM**. В частности, **LRM** имеет следующий вид: **Y ≈ -10.308*X + 5.232**.

Исследование данной модели показывает, что при фиксированном объеме (**n**) выборки изменение величины **M**-параметра в достаточно широких диапазонах (**M > 0**) не влияет на значения **CC** и **CR**-показателей. Тогда как при увеличении объема (**n**) выборки сохраняется прежнее соотношение между показателями **CC** и **CR** при увеличении степени связи между **Z**- и **X**-признаками (*рост значений* **CC** и **CR**), и уменьшении *средней квадратической ошибки* (**Er**) *выборочного* **CC**-показателя, как это иллюстрирует дальнейшее исследование данной модели регрессии (рис. **16**).

*Повтор начальной части Maple-документа из предыдущего фрагмента* (*рис*. **15**) *при условии замены значения* **100** *для* **n** *на* **1000**

```
> map(with, [linalg, SimpleStat]): `CC`:= CC(X, Z);
```

CC := -0.8204758882

```
> M:=matrix(2, 2, [SR(XX), SR(X), SR(X), 1]): B:=vector(2, [SR(ZX), SR(Z)]): r:=linsolve(M, B):
  Y:= r[1]*X + r[2]*E: u:= v -> r[1]*v + r[2]: u(v);
```

-10.27091317*v + 5.173084916

```
> DY:=sum((Z[k] - Y[k])^2, k=1..n)/n: `CR`:=sqrt((Ds(Z) - DY)/Ds(Z)); Er:=evalf((1 - CC(X,Z)^2)/
  sqrt(n-1)); n*(CR^2 - CC^2), abs(`CC`) - 3*Er;
```

CR := 0.8204758909

Er := 0.01034010559

0.44000*10^(-5), 0.7894555714

```
> Psi1, J:= 'sum((Z[k] - Y[k])^2/Y[k], k=1 .. n)', '(Psi1 - n)/sqrt(2*n)';
```

$$ш1, J := \sum_{k=1}^{n} \frac{(Z_k - Y_k)^2}{Y_k}, \quad \frac{ш1 - n}{\sqrt{2*n}}$$

```
> map(evalf, [Psi1, J]);
```

[728.8301827, -6.063541446]

**Рис. 16. Дальнейшая разработка модели регрессии примера (рис. 15). Проверка гипотезы о *линейном* характере *генерального* распределения на основе критерия согласия Романовского.**

На рис. **16** представлен *Maple*-документ, который в значительной степени дублирует начало предыдущего документа (рис. **15**) и позволяет вычислять основные *показатели корреляции* для *выборки* размера **n = 1000.** В результате выполнения данного документа получаем следующие результаты: **CC =** -0.8205, **CR =** 0.8205, **Er =** 0.01034 и некоторые другие. Дополнительно, на основе критерия согласия П. Романовского {*показатель* $\mathbf{J = (ш1 - n)/\sqrt{2*n}}$ [82, 92]} проверено соответствие в модели эмпирических **Z-**данных теоретическому **Y-**распределению. А именно, полученное значение **J = - 6.06354 << 3** говорит о существовании вполне *удовлетворительного* соответствия между **Z-**выборкой и **Y-**распределением, что в реальности не соответствует действительности. Аналогичный результат дает и критерий согласия Б.С. Ястремского [82, 91, 92, 155].

Таким образом, предложенная модель с учетом всего сказанного о методах вычисления связи между результативным и факторным признаками, однозначно иллюстрирует не только ненадежность **CC-** и **CR-**показателей, как измерителей степени связности признаков, но и самой базовой методологии, по крайней мере, *парного корреляционного анализа* в достаточно широком спектре приложений. Действительно, согласно им для нашей модели приемлема **ЛМР,** а в качестве достоверных измерителей связи **Z-** и **X-**признаков выступают **CC-** и **CR-**показатели, *совпадающие* по значению с высокой степенью точности. Тогда как *функциональная* зависимость между **Z** и **X** в *генеральной* совокупности имеет четко выраженный *периодический характер*, легко усматриваемый как из графического представления выборочных данных, так и самой линии регрессии (рис. **15, 16**).

Приведенная выше формальная *нелинейная* (*периодическая*) модель наглядно иллюстрирует тот факт, что в отрыве от качественного анализа чисто математический анализ проблемы связности явлений может приводить к недостоверным результатам, а именно: *традиционная*

*формальная методология корреляционного анализа вполне допускает исследование посредством* **ЛМР** *явлений при явно нелинейном характере связи между ними*; *при этом*, *все классические критерии допустимости такого типа моделей выполняются*. Из анализа приведеной периодической модели, наряду с сугубо *теоретическими*, можно сделать и несколько *полезных* практических и *методологических* выводов, а именно:

(**1**) *при проведении связного анализа явлений необходимо основную роль отвести логическо-качественному анализу их связи*, *базирующемуся на самой сущности явлений*; *формальный же анализ проводится относительно несложно с использованием стандартной методологии и средств* ***ВТ***

(**2**) *использование программных средств корреляционного анализа с развитыми графическими возможностями представления данных позволяет существенно облегчать задачу связного анализа*, *особенно с ее качественной стороны*; *существенную помощь здесь может оказать группировочный и другие методы связного анализа исследуемых явлений*

(**3**) *хорошее соответствие результатов связного анализа* ***ЛМР*** *при отсутствии должного логическо-качественного обеспечения может говорить о наличии*, *например*, *нелинейной связи периодического* (*циклического*) *характера между исследуемыми явлениями*

(**4**) *критерии согласия эмпирического и теоретического распределений не дают*, *в общем случае*, *высоко достоверных результатов*, *что требует их корректного применения*

(**5**) *метод наименьших квадратов не является наилучшёим методом построения теоретических распределений*; *однако*, *использование других методов требует*, *как правило*, *априорной информации о виде распределений*, *т.е. логическо-качественного предварительного анализа явлений*.

Учитывая приведенные и другие нечеткие аспекты *корреляционного* и *регрессионного* анализов, считаем целесообразным дополнительно проводить *связный* анализ методами компьютерной технологии, широко используя разнообразные графические средства отображения данных и возможности, предоставляемые современными статистическими прикладными пакетами, ориентированными, в первую очередь, на обширный класс **ПК.** *Компьютерный анализ связной* задачи и, возможно, некоторые другие простые приемы, близкие *по смыслу* к построению **CC**- и **CR**-*показателей*, смогут, на наш взгляд, значительно повысить достоверность результатов *связного анализа* при условии удовлетворительной их *логическо-качественной* интерпретации.

До последнего времени для построения экономико-статистических моделей применялись в основном методы *группировок* и методы *корреляционного* и *регрессионного* анализов. Между тем, насущная необходимость расширения формального аппарата экономико-статистического моделирования связана с объективными трудностями, которые продиктованы, прежде всего, невыполнением предпосылок использования *корреляционного* и *регрессионного* анализов, т. к. классические *теория вероятностей* и *математическая статистика* создавались применительно к анализу явлений природы. Социально-экономические же явления многомерны, дискретны, разнообразны, имеют случайную компоненту. Перечисленные особенности экономических процессов требуют применения в дополнение к аппарату классической статистики более универсальных методов математического описания. Одним из возможных путей решения данной проблемы является использование методов распознавания образов, как правило, на ЭВМ. Аппарат *теории распознавания образов* позволяет успешно выделять *однородные группы* по большому числу признаков, находить зависимости одновременно как от *количественных*, так от и *качественных* факторов. Методы *теории распознавания образов* можно применять почти на всех этапах экономико-статистического исследования, а именно: для выбора представителей групп, при анализе структуры выборочной совокупности, при обработке экспертных оценок и т.д.

# Глава 8.
# Элементы анализа временных рядов

В предыдущей главе рассматривались вариационные ряды, характеризующие вариацию признака некоторого явления безотносительно временной тенденции. Тогда как изучение *динамики* явления непосредственно связано с изменением значения того или иного признака со временем. Этой цели служат хронологические ряды, называемые также рядами *динамики* или *временными*. Методы *анализа временных* (*динамических*) рядов наиболее интенсивно начали развиваться в 20-х годах прошлого века и их развитие продолжается до сих пор [115, 123-125]. Процесс изменения явлений *во времени* заключается, главным образом, в том, что происходит изменение воздействия на них многих факторов, и само время становится собирательным фактором – внешне явление развивается под действием времени. Анализ временных рядов является важной экономико-статистической задачей, позволяющей изучать разнообразные явления в их *динамике*, определять их *тенденции* и разрабатывать соответствующие *прогнозы*.

Более формально, *временной ряд* может быть определен следующим образом. ***Временной ряд*** – *набор данных*, который устанавливает значения *результирующей переменной* и соответствующих значений *времени* или связанной со временем переменной по *временному интервалу*. Если же данные включают значения *результирующей* **Y**-*переменной*, зарегистрированной в *дискретные* моменты времени, например **(Y1, t1), (Y2, t2), …, (Yn, tn),** то ряд, говорят, является ***дискретным*** *временным рядом*. Как правило, *моменты*, в которые производится *регистрация*, являются *равно удаленными*. Например, рис. **17** – графическое представление дискретного временного ряда с равно удаленными точками отсчета времени. В дальнейшем наряду с термином "*временной ряд*" будем использовать его синонимы "*динамический ряд*" и просто "*ряд*"**.**

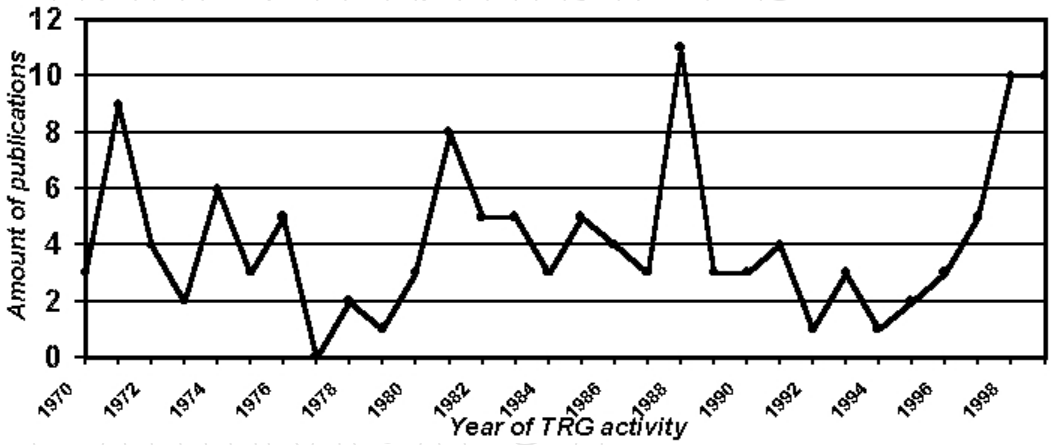



**Рис. 17. График дискретного временного ряда, базирующегося на данных табл. 9 (*графа* 6)**

Данный *дискретный* временной ряд представляет *ежегодную динамику* количества публикаций **TTG** на основе данных графы **6** из табл. **9.** Альтернативно, *непрерывный* временной ряд может быть определен, в котором значения **Y** определены для *каждого значения величины* **t** в заданном временном интервале. Практически, *непрерывные* временные ряды возникают в

случае, когда: (**1**) *устройство регистрации фиксирует значения переменной по временному интервалу и делает запись результатов в виде непрерывного графика* или (**2**) *результирующая переменная дискретна и моменты, в которых изменения имеют место, отмечаются*. Обратите внимание, что термины *"непрерывные"* и *"дискретные"* в применении к *временному* (*динамическому*) ряду не *определяют* тип результирующей переменной. Как *дискретный*, так и *непрерывный* временной ряд может явиться результатом либо дискретных, либо непрерывных результирующих переменных. Непрерывные временные ряды имеют достаточно ограниченное и специализированное применение, и не рассматриваются далее в настоящей книге. Методы анализа *дискретного* временного ряда наиболее интенсивно начали развиваться в двадцатых годах позапрошлого столетия, и их развитие продолжается до настоящего времени [115, 123-125, 127, 132, 157, 182, 214, 215, 270-272, 285].

Между тем, временные ряды не присущи только экономическим и статистическим разделам. Они играют достаточно важную роль, например, в такой междисциплинарной науке как теория систем, когда *динамическая система* представляется неким объектом, состоящим из некоторого семейства временных рядов [214]. На основе исследования методов динамических систем может успешно быть решена формализация проблемы моделирования временного ряда. Данное исследование важно как на *концептуальном*, так и на *алгоритмических* уровнях. Алгоритмы приблизительного анализа моделирования временного ряда представляют непосредственный интерес в многочисленных приложениях, например, в эконометрике. В частности, интересный системный подход к разработке точных и приблизительных моделей, объясняющих заданный (*наблюдаемый*) временной ряд может быть найден в работе [214].

## 8.1. Типы временных рядов, их построение и представление

Ряд расположенных во временной шкале значений статистического показателя, изменение которых отражает закономерность развития изучаемого явления, называется ***временным*** (*динамическим*). Каждый *временной ряд* состоит из двух компонент, а именно: (**1**) ***временная шкала*** – моменты (*даты, времена и др.*) или периоды (*месяц, квартал, год, пятилетка и др.*) времени; (**2**) ***уровни ряда*** – собственно статистические данные, относящиеся к выбранной временной шкале. Обе компоненты называются членами временного ряда. Особо выделяют уровни ***начальный, конечный*** и ***средний,*** представляющие собой соответственно *первый, последний* уровни временного ряда и среднюю из уровней ряда. Уровни временного ряда имеют две основные особенности, а именно: (**1**) *уровень последующего времени зависит от уровня предыдущего времени* и (**2**) *чем больше временной интервал между событиями, тем больше, как правило, различаются их количественные и качественные характеристики*. Например, в табл. **4** представлена динамика по пятилеткам научных публикаций **ТТГ,** из которой видно, что данный временной ряд подвержен существенным колебаниям относительно средней арифметической. Однако по другому признаку – качество публикаций, определяемому их цитируемостью, временной ряд (табл. **9**) указывает на постоянный рост интереса к работам **ТТГ,** в результате чего группа занимает одну из ведущих позиций среди исследовательских коллективов по **МТОС** и ее многочисленным приложениям, особенно в вычислительных и биологических науках.

Для правильного анализа временного ряда необходимо знать их типы, определяемые при группировке элементов ряда по разным признакам. По *временной шкале* они делятся на ***моментные*** и ***интервальные***. В *моментном* временном ряде его уровни выражают величину явления на определенные дату или время. Примером такого временного ряда может служить табл. **13,** в которой дана *динамика* суммарного количества крупных публикаций (*монографии, книги, сборники статей, отчеты*) **ТТГ** на конец завершающего каждую пятилетку года. Не

имеет смысла суммировать *уровни* мгновенного временного ряда, но ***разность уровней*** ряда имеет весьма определенный смысл.

**Таблица 13. Суммарные количества крупных публикаций ТТГ**

| *Итоговая дата* | 32.12.74 | 31.12.79 | 31.12.84 | 31.12.89 | 31.12.94 | 31.12.99 |
|---|---|---|---|---|---|---|
| *Публикация* | 2 | 5 | 11 | 22 | 36 | 53 |

В ***интервальном*** ряде его уровни выражают величину явления за определенный период времени (*год, пятилетку и др.*). Отличительной особенностью *интервального* ряда *абсолютных* величин является возможность суммирования уровней смежных интервалов, укрупняя их. В результате получаем накопленные (*кумулятивные*) итоги, имеющие реальное содержание. Следует, однако, иметь в виду,что в ряде случаев возможно преобразование *моментного* ряда в *интервальный*, и наоборот. Примером интервального временного ряда служат данные табл. **7,** содержащей распределение публикаций **ТТГ** по годам ее деятельности. В данной таблице можно выделить не менее 8 временных рядов – динамики публикаций **ТТГ** за определенный период (*год*) в разрезе их (**1**) типа (*книги*, *статьи*), (**2**) места издания (*СССР*, *за рубежом*), (**3**) объема в страницах оригинального издания и т.д.

По полноте *временного* представления ряды делятся на ***полные*** и ***неполные***. В *полных* рядах временная шкала имеет равные интервалы, в *неполных* – неравные (*уровни по некоторым моментам или интервалам отсутствуют*). Приведенные выше временные ряды (табл. **7, 13**) являются *полными*. По способу выражения уровней ряды делятся на ряды ***абсолютных, средних*** и ***относительных*** величин. Приведенные до сих пор примеры относились к рядам первого типа. В табл. **14** представлена динамика среднего объема крупных публикаций (*монографии*, *книги*, *сборники статей*, *отчеты*) по пятилеткам деятельности **ТТГ**.

**Таблица 14. Динамика средних объемов (*в стр.*) монографических публикаций ТТГ**

| *Период* | 1970 - 1974 | 1975 - 1979 | 1980 - 1984 | 1985 - 1989 | 1990 - 1994 | 1995 - 1999 |
|---|---|---|---|---|---|---|
| *Средний объем* | 71 | 118 | 208 | 370 | 659 | 1334 |

Данные табл. **14** легко получаются из табл. **7** (графа **3**); из них, в частности, следует, что динамика средних размеров монографических публикаций **ТТГ** имеет ярко выраженную тенденцию к росту, иллюстрируемому следующим графическим (рис. **18**) представлением ряда средних объемов (*average sizes*).

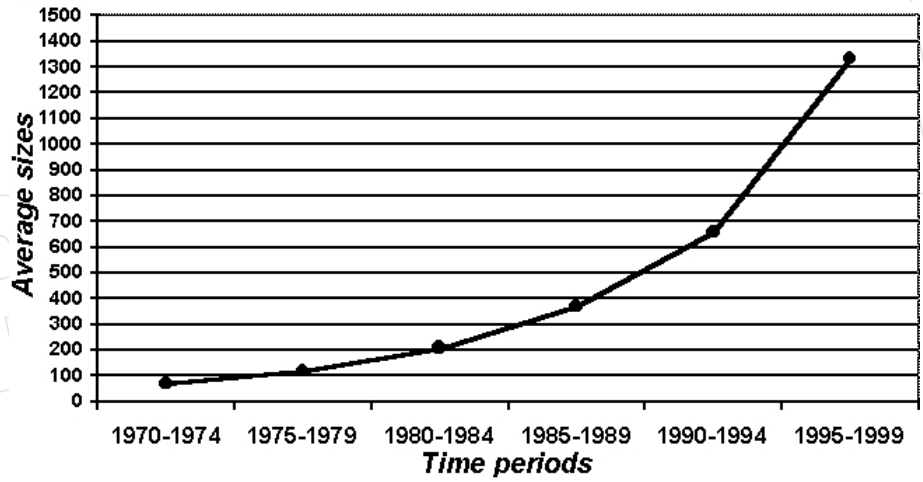



**Рис. 18. Динамика средних объемов публикаций ТТГ**

Примером временного ряда *относительных* величин может служить табл. **5,** отражающая динамику по пятилеткам доли публикаций **ТТГ** по их типам (*монографии*, *книги*, *статьи*, *сборники*, *научные отчеты*).

Важнейшей проблемой построения временных рядов является *сопоставимость* их уровней, ибо при ее отсутствии невозможно получать достоверных показателей динамики рядов. Приведем основные причины несопоставимости уровней временного ряда и способы их устранения. Важным требованием любых динамических сравнений является сопоставимость территории, к которой относятся уровни ряда. Изменение территории обычно приводит к различию статистических показателей. Для приведенияэтих данных к сравнимому виду производится их перерасчет с учетом новых территориальных границ. Например, средняя величина миграции населения в СССР была одна, а с его распадом страны эта величина существенно изменилась, например, для СНГ и ее субъектов. Понятие территории носит и более широкий смысл. Так, научная активность одного и того же творческого коллектива неодинакова в различных местах его дислоцирования, что проверено на практике.

*Уровни* ряда должны быть сопоставимы по *кругу охватываемых объектов*. Несопоставимость может возникать, например, вследствии переподчинения составляющих объект единиц. Однако сопоставимость не нарушается при уменьшении/увеличении единиц совокупности. Примером несопоставимости может служить совмещение в одной совокупности публикаций монографического и периодического характеров. Вполне очевидно, их динамики в процессе научной активности, в общем случае, различны. Сопоставимость по кругу охватываемых объектов достигается путем перехода к относительным показателям и смыканием рядов. Приведем пример смыкания временных рядов объемов услуг фирмы *SALCOMBE Eesti Ltd*., к которой в 1993 была присоединена фирма *FIDO Ltd*. Реализация услуг объединенной фирмы выражается следующим временным рядом (см. табл. **15**).

**Таблица 15. Объем оказанных информационных услуг фирмой *SALCOMBE Ltd* (*в тыс. EEK*)**

| ***Период*** | **90** | **91** | **92** | **93** | **94** | **95** | **96** | **97** | **98** | **99** |
|---|---|---|---|---|---|---|---|---|---|---|
| ***Перед слиянием*** | **120** | **130** | **135** | **140** | | | | | | |
| ***После 1-го слияния*** | | | | **170** | **175** | **180** | **190** | | | |
| ***После 2-го слияния*** | | | | | | | **235** | **250** | **260** | **300** |

В самом простом случае слияния двух фирм, как правило, *год слияния* выбирается в качестве ***базисного*** *года* и относительно этого года делается перевычисление. Однако, в нашем случае имеют место два слияния в различные годы (**1993** и **1996**). Поэтому, целесообразно выбрать в качестве ***базисного*** года год, расположенный между годами слияния. Без потери общности мы предполагаем, что таким годом является **1995.** Для закрытия *временного* ряда, *отражающего* динамику объемов услуг *SALCOMBE Eesti Ltd*, мы принимаем данные **1995** (***базисный*** *год*) за 100% относительно предыдущих и последующих лет (табл. **15**). Поэтому, для объединенной *SALCOMBE Eesti Ltd* в 1995 выбирается показатель **180**. Произведя простые вычисления, мы получаем закрытый временной ряд (табл. **16**) объемов предоставленных услуг (*в процентах к базисному* **1995**). Точно так же мы выполняем закрытия ряда относительно базисных лет (**1993** и **1996**), которые одновременно являются и годами слияния с фирмой *SALCOMBE Eesti Ltd* вышеупомянутых двух фирм. Однако, в случае выбора года слияния в качестве ***базисного,*** *перевычисление* производится несколько иначе, а именно. Без потери общности, мы выбираем **1996** в качестве *базисного* года. При сделанном предположении, за 100% для предыдущих лет выбирается значение **190,** а для последующих – значение **235** соответственно (табл. **15**). Итак, произведя (*принимая во внимание вышеупомянутые соображения*) простые вычисления, мы

получаем закрытый временной ряд (табл. **16**) объемов оказанных информационных услуг (*в процентах к базисному* **1996**). Аналогично мы решаем данную задачу и для случая базисного года **1993** (табл. **15**). Итоговые результаты *закрытия* временного ряда относительно указанных базисных лет представлены в табл. **16.**

**Таблица 16. Динамика объемов оказанных информационных услуг фирмой *SALCOMBE Ltd***

| *Период* | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 |
|---|---|---|---|---|---|---|---|---|---|---|
| *Относительный уровень в % к 1993* | 85.7 | 92.9 | 96.4 | 100 | 102.9 | 105.9 | 138.2 | 147.1 | 152.9 | 176.5 |
| *Относительный уровень в % к 1995* | 66.7 | 72.2 | 75.0 | 94.9 | 97.2 | 100 | 130.6 | 138.9 | 144.4 | 166.7 |
| *Относительный уровень в % к 1996* | 63.2 | 68.4 | 71.0 | 89.5 | 92.1 | 94.7 | 100 | 106.4 | 110.6 | 127.7 |

На основе *графического* представления (рис. **19**) полученного временного ряда *относительных* величин, значения которых представлены табл. **16,** мы имеем возможность удостовериться, что динамика ряда (*прямая на рис*. **19** *отмечена красным*), полученная *закрытием* относительно *базисного года* **1995,** который является *промежуточным* между годами **1993** и **1996** *слияния* фирм, отражает среднюю динамику временного ряда, полученного закрытием относительно лет слияния. Сглаживая *критические* годы слияния, приведенный метод во многих случаях более предпочтителен при разработке некоторого временного ряда *относительных* величин. При этом, успешность *закрытия* ряда *относительных* величин в значительной степени зависит от выбранного метода, который должен принимать во внимание характер *явления*, отражаемого искомым временным рядом.

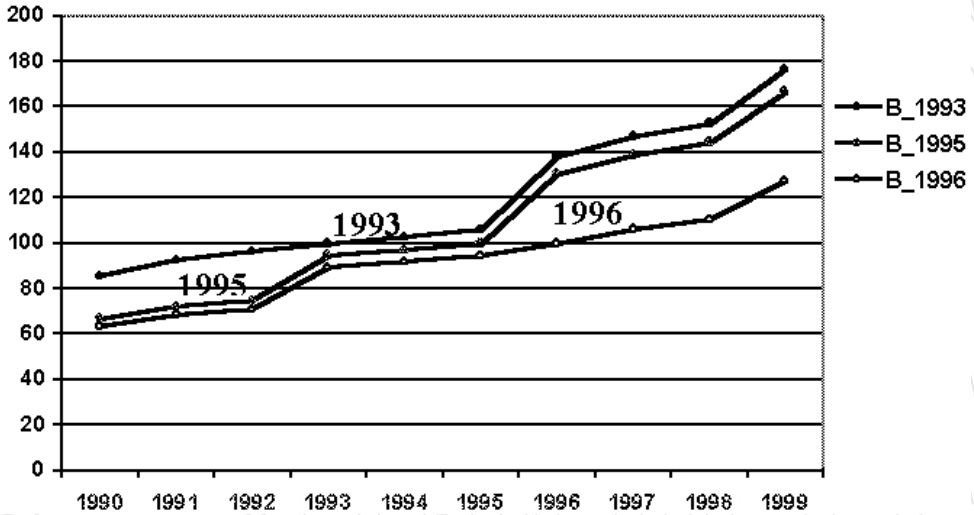



**Рис. 19. Динамика объемов оказанных услуг фирмой в зависимости от *базисного* года**

В ***моментных*** временных рядах может возникать *несопоставимость* по *критическому моменту регистрации* для явлений с сезонным характером уровней. Например, число заболеваний ОРЗ выше зимой и ранней весной, чем летом. Поэтому, нельзя в такие ряды включать уровни с разными датами регистрации. Несопоставимость из-за различия *единиц измерений* очевидна сама по себе. Она может возникнуть, например, из-за несравнимости денежных оценок, что особенно актуально в условиях постсоветского периода (*оптовые*, *розничные и региональные*

*цены*, *различные денежные единицы и др.*). Для устранения такого рода несопоставимости можно использовать различные индексы. Уровни временного ряда должны быть сопоставимы и по методике их расчета, т.е. при их сведении в ряд следует оценивать методики их получения на предмет совместимости.

Условием сопоставимости уровней *интервального* ряда является ***равенство периодов времени,*** за которые приводятся данные. Несопоставимость статистических данных может возникнуть из-за *различного понимания* единиц совокупности, характеризуемой исследуемым временным рядом. Например, сведение в один временной ряд уровней (*объемы публикаций*) для статей и книг заведомо не является корректным по причине как их периодичности изданий, так и большой разнице в объемах. Итак, исследуемая совокупность должна быть однородной по выбранному признаку, значения которого и составляют уровни ряда. Несопоставимость статистических показателей динамики может быть обусловлена и различной структурой совокупности за различные периоды времени. В связи со сказанным, вопросу *сопоставимости* уровней временного ряда в каждом конкретном случае следует уделять самое пристальное внимание во избежание ошибочных результатов анализа динамики исследуемого явления.

***Графическое представление*** динамики уровней временного ряда позволяет весьма наглядно изображать процесс развития явления во времени, упрощает многие вопросы последующего сглаживания рядов, а также весьма способствует более эффективному их анализу. Способы графического представления (*подробнее о них говорилось в разделе* **4.6;** *см. также рис.* **17, 18, 19**) динамики рядов весьма разнообразны, здесь рассмотрим лишь некоторые из них. Наиболее широко используемым является представление ряда ***линейной диаграммой,*** которая строится в декартовой системе координат – на осях абсцисс и ординат откладываются сответственно *даты* или *периоды времени* и *уровни* ряда. При представлении *интервального* временного ряда его уровни относятся к *серединам* соответствующих им интервалов.

Рассмотрим графическое представление *интервальных* рядов **U, A** и **G,** характеризующих динамику цитирования в СССР и за рубежом публикаций **ТТГ** (табл. **9;** столбцы **4, 5** и **6** соответственно), в единых координатах (рис. **20**). Изображение в одной системе координат нескольких графиков различными типами линий позволяет легко визуально сопоставлять динамику разных показателей. Но в случае большого числа графиков (*как правило, более* **4**) может проигрывать наглядность представления. Рис. **20** показывает явную тенденцию к *увеличению* цитируемости публикаций **ТТГ** как в СССР, так и за рубежом при существенном опережении во втором случае, что может в определенной мере характеризовать уровень развития **МТОС** и ее приложений в СССР и за рубежом. Данные графики представлены в единой системе координат посредством достаточно известного средства "*Diagram Microsoft Graph 97*" из известного пакета "*Microsoft Office 97*" для *Windows 98SE*.

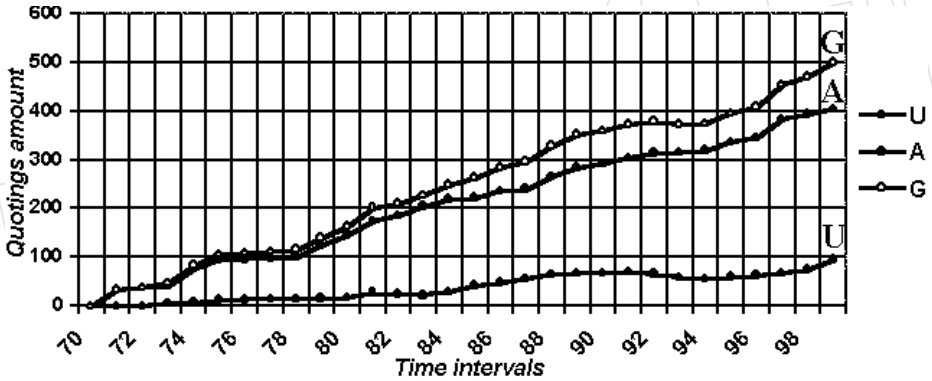



**Рис. 20. Графическое представление трех *интервальных* временных рядов, отражающих ежегодную динамику цитирования публикаций ТТГ относительно их места издания**

Статистические кривые временных рядов различаются по *типу* используемых координат, точнее по масштабам осей координат, что приводит к изменению конфигурации кривых. Чаще всего используются *равномерные* масштабы: по осям абсцисс и ординат отрезки шкалы берутся пропорционально числу соответственно периодов времени или дат, и также самим уровням. По оси ординат может использоваться шкала, пропорциональная либо самим уровням, либо их логарифмам. Этот тип координат используется для временных рядов с резко изменяющимися уровнями. Однако, он придает статистическим кривым слабую тенденцию вверх и как бы скрадывает динамику самого процесса. Такая система координат называется *полулогарифмической* и широко используется при изображении динамики ряда важных экономических показателей.

Кроме *линейных* диаграмм могут применяться *радиальные*, весьма удобные для представления динамики процессов *периодического* (*например*, *сезонного*) характера. В радиальной системе координат на окружности откладывается временная шкала, а на ее диаметрах уровни ряда, начиная от центра. Получаемые концентрические ломаные изображают ряд с повторением его циклических колебаний. Однако, неудачный выбор соотношения масштабов по осям координат может существенно изменить внешний вид статистической кривой временного ряда и создать неверное представление о *динамике* изучаемого явления, представляемого им.

Для сравнения отдельных уровней временного ряда используются *столбиковые* диаграммы; при этом, столбики уровней можно располагать вплотную и раздельно. Например, на рис. **21** представлена *столбиковая диаграмма* динамики *отечественных* (*Domestic*) и *зарубежных* (*Foreign*) публикаций **ТТГ** в области **МТОС** и ее приложений по пятилеткам активности. Графическое представление ряда позволяет визуально определять основные черты динамики изучаемого явления или сравнительной динамики нескольких явлений, или различных сторон одного и того же явления. В последнем случае рекомендуем придерживаться *единой* временной шкалы координат либо использовать *относительные* величины, когда статистические *кривые сходятся* в *одной* (*базисной*) точке, что позволит легко проводить различные *сопоставления* исследуемых временных рядов.

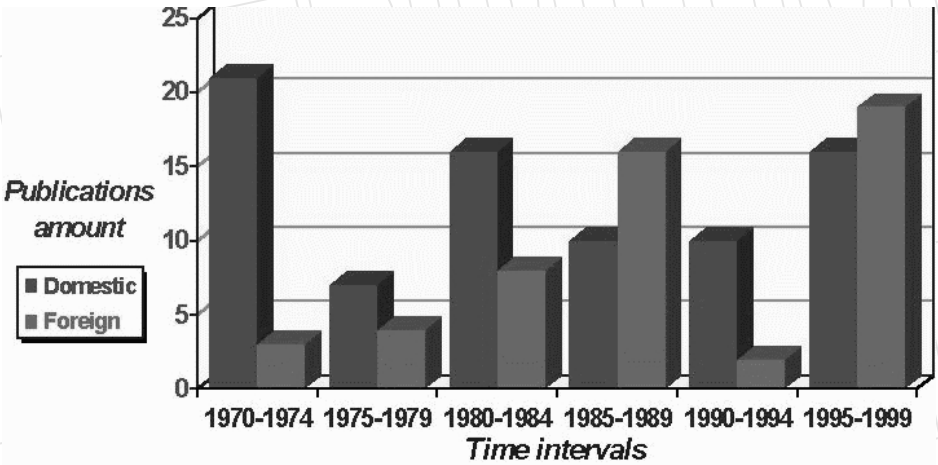



**Рис. 21. Количества отечественных и зарубежных публикаций ТТГ по пятилеткам**

Необходимые данные для разработки и построения вышеуказанных диаграмм получены на основе статистических данных наблюдения активности **ТТГ** (табл. **9;** графы **2** и **3**).

## 8.2. Статистические показатели временного ряда

Для характеристики ***особенностей*** и ***закономерностей*** развития изучаемого явления во времени необходимо решить целый ряд задач и осветить широкий круг вопросов. К числу

основных задач, возникающих при изучении временных рядов, относятся следующие: (**1**) *характеристика интенсивности отдельных изменений в уровнях ряда с течением времени*; (**2**) *определение средних показателей динамики*; (**3**) *выявление закономерностей динамики ряда в целом*; (**4**) *интерполяция и экстраполяция*; (**5**) *выявление основных факторов, влияющих на изменение динамики исследуемого явления*.

***Временной ряд*** – это *ряд последовательных уровней* (**L**), сопоставляя которые между собой, можно получать характеристику скорости и интенсивности развития явления. В результате сравнения уровней получается ***система*** *абсолютных* и *относительных* показателей динамики, к числу которых относятся: *абсолютный прирост*, *коэффициент* (*темп*) *роста*, *темп прироста*, *абсолютное значение* **1%** и др. При этом, сравниваемый уровень ряда называется ***текущим,*** а уровень, с которым производится сравнение – ***базисным***. За *базисный* часто принимается либо *начальный* в ряде уровень, либо уровень, с которого начинается какой-то новый этап развития явления, отражаемого данным рядом. Если производится сравнение каждого уровня ряда с предыдущим, то имеем ***цепные показатели*** динамики; при сравнении с базисным уровнем – ***базисные показатели*** динамики. В первом случае показатели динамики ряда характеризуют интенсивность изменения уровней от периода (*момента*) к периоду (*моменту*), во втором – окончательный результат всех изменений в уровнях ряда за период от базового уровня до текущего. Выбор базиса должен быть обоснованным и отвечать самой сущности изучаемого явления.

***Абсолютный прирост*** (**AIC**) есть разность двух смежных уровней ряда, т.е.: $\mathbf{AIC_k = L_k - Lb}$, где $\mathbf{L_k}$ и **Lb** – соответственно *текущий* и *базисный* уровни ряда ($\mathbf{k = 1 .. n}$). Иногда показатель **AIC** называют ***скоростью роста.*** В качестве *базисного* уровня можно выбирать как некоторый фиксированный (**Lb =** *constant*, *basis*), так и предыдущий ($\mathbf{Lb = L_{k-1}}$) уровень. Как правило, за базисный выбирается первый ($\mathbf{Lb = L_1}$) уровень ряда. Не нарушая общности, в дальнейшем будем полагать ***базисным*** *первый уровень* временного ряда. За весь период, описываемый временным рядом, величина **AIC** вычисляется по следующей простой формуле:

$$\mathbf{AIC} = \sum_{k=1}^{n} \mathbf{AIC_k} = \mathbf{L_n - L_1}, \qquad \textit{if} \quad \mathbf{L_b \equiv L_{k-1}}$$

Для иллюстрации расчетов показателей динамики временного ряда определяем ряд (табл. **17;** графы **1** и **2**) на основе табл. **9** (графа **3** – *цитируемость публикаций* **ТТГ** *в СССР по* **МТОС**).

**Таблица 17. Динамика годовой цитируемости публикаций ТТГ в СССР**

| *Год* | Lk | AICk | GCk | GCbk | GRk | GRbk | A1%k |
|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) |
| 1972 | 0 | 0 | - | - | - | - | - |
| 1973 | 5 | 5 | 1.00 | 1.00 | 0.00 | 0.00 | - |
| 1974 | 7 | 2 | 1.40 | 1.40 | 0.40 | 0.40 | 0.05 |
| 1975 | 12 | 5 | 1.71 | 2.40 | 0.71 | 1.40 | 0.07 |
| 1976 | 13 | 1 | 1.08 | 2.60 | 0.08 | 1.60 | 0.12 |
| 1977 | 14 | 1 | 1.08 | 2.80 | 0.08 | 1.80 | 0.13 |
| 1978 | 14 | 0 | 1.00 | 2.80 | 0 | 1.80 | 0.14 |
| 1979 | 15 | 1 | 1.07 | 3.00 | 0.07 | 2.00 | 0.14 |
| 1980 | 18 | 3 | 1.20 | 3.60 | 0.20 | 2.60 | 0.15 |
| 1981 | 27 | 9 | 1.50 | 5.40 | 0.50 | 4.40 | 0.18 |
| 1982 | 24 | -3 | 0.89 | 4.80 | -0.11 | 3.80 | 0.27 |
| 1983 | 22 | -2 | 0.92 | 4.40 | -0.08 | 3.40 | 0.24 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **1984** | **29** | **7** | **1.32** | **5.80** | **0.32** | **4.80** | **0.22** |
| **1985** | **42** | **13** | **1.45** | **8.40** | **0.45** | **7.80** | **0.29** |
| **1986** | **48** | **6** | **1.14** | **9.60** | **0.14** | **8.60** | **0.42** |
| **1987** | **56** | **8** | **1.17** | **11.20** | **0.17** | **10.20** | **0.48** |
| **1988** | **64** | **8** | **1.14** | **12.80** | **0.14** | **11.80** | **0.56** |
| **1989** | **68** | **4** | **1.06** | **13.60** | **0.06** | **12.60** | **0.64** |
| **1990** | **68** | **0** | **1** | **13.60** | **0** | **12.60** | **0.68** |
| **1991** | **70** | **2** | **1.03** | **14.00** | **0.03** | **13.00** | **0.68** |
| **1992** | **65** | **-5** | **0.93** | **13.00** | **-0.07** | **12.00** | **0.70** |
| **1993** | **58** | **-7** | **0.89** | **11.60** | **-0.11** | **10.60** | **0.65** |
| **1994** | **55** | **-3** | **0.95** | **11.00** | **-0.05** | **10.00** | **0.58** |
| **1995** | **60** | **5** | **1.09** | **12.00** | **0.09** | **11.00** | **0.55** |
| **1996** | **63** | **3** | **1.05** | **12.60** | **0.05** | **11.60** | **0.60** |
| **1997** | **67** | **4** | **1.06** | **13.40** | **0.06** | **12.40** | **0.63** |
| **1998** | **75** | **8** | **1.12** | **15.00** | **0.12** | **14.00** | **0.67** |
| **1999** | **95** | **20** | **1.27** | **19.00** | **0.27** | **18.00** | **0.75** |
| **Итого** | **1154** | **95** | **-** | **-** | **-** | **-** | **-** |

Для случая нашего примера получаем значение **AIC** = (95-0) = 95. Абсолютные уровни $\mathbf{L_k}$ этого ряда, исключая годы 1982, 1983, 1990 и 1992 - 1994, непрерывно растут до 1999 г.; данный факт отражает и поведение величин $\mathbf{AIC_k}$ абсолютного прироста.

Кроме того, следует сказать несколько слов о выборе *базисного уровня* для нашего конкретного временного ряда. Прежде всего, публикации ТТГ восходят к 1970 г.; однако, регистрация их цитирования была начата только в 1971. При этом, из-за совершенно понятных причин в начале регистрации были получены нулевые значения для отечественного цитирования (*неполнота охвата рассмотренных материалов*, *естественная задержка времени с момента начала публикации до первых ссылок на нее и*, *прежде всего*, *недостаток в тот момент серьезного интереса к данной проблематике со стороны ведущих советских исследователей*). В свете вышесказанного, в качестве *первого* уровня временного ряда для *отечественного* цитирования был выбран 1972 г., а в качестве *базисного* уровня – 1973, тогда как в случае *зарубежного* цитирования желательно выбрать 1970 и 1971 в качестве *первого* и *базисного* уровней соответственно (табл. **9;** графа **5**).

***Коэффициент роста*** (**GC**) – отношение *текущего* уровня к *предыдущему* или к некоторому *базисному***;** он определяет *темп роста*. Величина показателя **GC** может быть безразмерной или выражаться в процентах; при этом, сам уровень ряда, с которым производится сравнение, принимается за 100%**.** Темп роста характеризуется коэффициентами: ***цепными*** ($\mathbf{GC_k = L_k / L_{k-1}}$), ***базисными*** ($\mathbf{GCb_k = Lk/Lb}$) и ***абсолютными*** ($\mathbf{GC = L_n/L_1}$). Величина **GC** *коэффициента роста* показывает для $\mathbf{GC > 1}$, $\mathbf{GC < 1}$ или $\mathbf{GC = 1}$ соответственно *увеличение*, уменьшение или неизменность *темпа роста* текущего уровня относительно *сравниваемого*. Если *коэффициенты роста* выражаются в процентах, то они называются ***темпами роста***. Для примера из табл. **17** получаем значение **GC** = (95/5) = 19**;** значения для $\mathbf{GC_k}$ и $\mathbf{GCb_k}$ (*в долях*) приводятся в графах **3** и **4** (*базисным принят уровень* 1973 *г.*).

***Темпы прироста*** (**GR**) временного ряда вычисляются по следующим простым формулам: $\mathbf{GR_k = AIC_k/L_{k-1} = (L_k/L_{k-1}) - 1 = GC_k - 1}$ and $\mathbf{GRb_k = AIC_k/Lb = (L_k/Lb) - 1 = GRb_k - 1}$ ($\mathbf{k = 1 .. n}$). Они показывают на сколько увеличился или уменьшился *текущий* уровень относительно *предыдущего* или *базисного* уровней. Темпы прироста могут быть как безразмерными коэффициентами, так и выражаться в процентах. Так как *абсолютный прирост* за весь период ряда равен $\mathbf{AIC = L_n - L_1}$**,** то за этот же период *темп прироста* составит уже величину **GR =**

**AIC / $L_1$ = GC - 1.** Для случая примера табл. **17** получаем **GR** = 19 - 1 = 18; величины коэффициентов $GR_k$ и $GRb_k$ приведены соответственно в графах **6** и **7** табл. **17.**

Наконец, ***абсолютное значение* 1% *прироста*** (**A1%**) показывает, какая абсолютная величина скрывается за относительным показателем – одним процентом прироста. Данный показатель вычисляется по следующим простым формулам: **A1% = AIC / (GR*100) = 0.01*Lb** и $A1\%_k$ = $AIC_k / (GR_k*100) = 0.01*L_{k-1}$. Он представляет собой отношение абсолютного прироста к темпу прироста, выраженному в процентах (%). Простое преобразование показывает, что величина **A1%** есть не что иное, как сотая часть *базисного уровня* **Lb,** в качестве которого может быть либо *предыдущий* уровень ($L_{k-1}$) ряда, либо некоторый *фиксированный* (*например*, $L_1$). В случае нашего примера (табл. **17**) оба показателя **GC** (графа **4**) и **A1%** (графа **8**) имеют весьма существенно различные тенденции динамики, тсключая последнюю пятилетку (1995 – 1999), где их тенденции, практически, идентичны. Рис. **22** хорошо иллюстрирует данный факт.

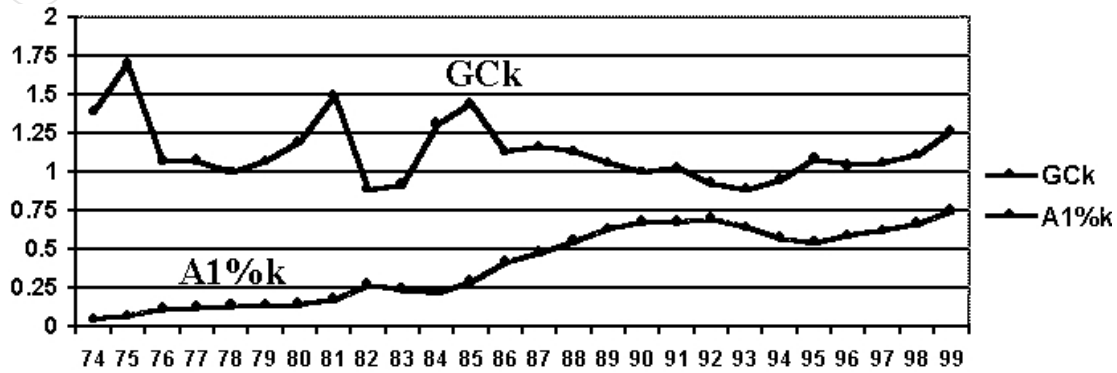



**Рис. 22. Динамика показателей GC и A1% для временного ряда из табл. 17**

Между показателями динамики, вычисленными с *постоянной* (**Lb**) и *переменной* ($L_{k-1}$) базой, существует простая связь, установление которой (*часть данной работы уже проделана выше*) в формульном виде предоставляется читателю в качестве полезного упражнения. В случае сопоставления динамики развития двух явлений можно использовать ***коэффициент*** ($LC_k$) ***опережения,*** вычисляемый по следующей простой формуле:

$$LC_k = \frac{L^1_k * L^2_1}{L^2_k * L^1_1}$$

где $L^1_k$ и $L^2_k$ – соответственно **к**-й уровень *первого* и *второго* временного ряда. Показатель $LC_k$ представляет собой отношение темпов роста за одинаковые отрезки времени по двум рядам. С помощью данного показателя могут сравниваться временные ряды *одинакового* содержания, но относящиеся к разным объектам, или ряды *разного* содержания, но относящиеся к одному и тому же объекту. В качестве полезного упражнения читателю рекомендуется вычислить коэффициенты опережения для двух рядов, описывающих динамику цитируемости работ **ТТГ** по **МТОС** и ее приложениям в СССР и за рубежом (табл. **9;** графы **4** и **5** соответственно), и дать им содержательную интерпретацию, которая сможет пролить свет на природу взаимосвязи между обоими временными рядами.

## 8.3. Средние показатели временного ряда

Для получения *обобщающей* характеристики динамики исследуемого явления определяются различного рода *средние* показатели. Мы рассмотрим только *две* категории таких показателей: *средние уровни* и *средние изменения уровней* временного ряда. Для *интервального временного* ряда

*абсолютных* показателей ***средний уровень*** ($\bar{L}_{int}$) за весь период временного ряда определяется по формуле простой средней арифметической, а именно:

$$\bar{L}_{int} = \frac{\sum_{k=1}^{n} L_k}{n} \tag{65}$$

где $L_k$ – уровни ряда и $n$ – их количество ($k = 1 .. n$). В случае примера табл. **17** (графы **1** и **2**) получаем значение *среднего уровня* $\bar{L}_{int} = 1154/28 = 41.21$.

***Средний уровень*** ($\bar{L}_{ins}$) *моментного* временного ряда определяется несколько иначе. Пусть $L = \{L_1, L_2, ..., L_n\}$ – *моментный* ряд с равными временными промежутками между датами. Мысленно заменяем его *интервальным* [*с* (**n-1**) *интервалами*] рядом в предположении, что уровень между смежными датами (*на интервале*) изменяется равномерно и непрерывно. Тогда значение уровня на интервале воображаемого ряда можно определить как $L_k = (L_k + L_{k+1})/2$ при условии ($k = 1 .. n-1$), тогда как *его средний уровень* вычисляется по формуле:

$$\bar{L}_{ins} = \frac{\sum_{k=1}^{n-1} L_k}{n-1} + \frac{L_n - L_1}{2*(n-1)} \tag{66}$$

т.е. ***средний уровень*** *моментного* ряда равен сумме средней арифметической первых его (**n-1**) уровней и полуразности его крайних уровней, деленной на число дат без одной. Вычисление среднего уровня моментного ряда, определяемого табл. **13,** дает результат $\bar{L}_{ins} = (2 + 5 + 11 + 22+36)/5+(53-2)/(2*5)=20.3$, а не значение 21.5, если бы величина $\bar{L}_{ins}$ вычислялась по *формуле* простой средней арифметической.

Если имеются два ряда – *интервальный* и *моментный* с одним и тем же множеством **m** уровней [$\#(M) = n$], то их *средние значения уровней* связаны следующим линейным соотношением:

$$\bar{L}_{int} = \frac{n-1}{n} * \bar{L}_{ins} + \frac{L_1 + L_n}{2*n} \tag{67}$$

которое нетрудно получить на основе формул (65, 66). *Из* формулы (67) путем несложных геометрических соображений можно получить следующие практически полезные оценки:

$$\textbf{(1)} \quad \bar{L}_{int} > \bar{L}_{ins} \quad \textit{if} \quad \bar{L}_{int} < (L_1+L_n)/2$$
$$\textbf{(2)} \quad \bar{L}_{int} < \bar{L}_{ins} \quad \textit{if} \quad \bar{L}_{int} > (L_1+L_n)/2$$
$$\textbf{(3)} \quad \bar{L}_{int} = \bar{L}_{ins} \quad \textit{if} \quad \bar{L}_{int} = (L_1+L_n)/2$$

Во всех приведенных расчетах средних уровней предполагалось, что временные ряды были *полными*. При *неполных уровнях* используется взвешивание сумм *каждой смежной* пары уровней по продолжительности периода времени между ними, а именно:

$$\bar{L}_{ins} = \frac{\sum_{k=1}^{n-1} (L_k + L_{k+1}) * T_k}{2 * \sum_{k=1}^{n-1} T_k}$$

где $T_k$ – время между моментами уровней $L_k$ и $L_{k+1}$ *моментного* ряда. В знаменателе берется удвоенная сумма периодов, ибо каждое слагаемое числителя суммируется дважды.

***Средний абсолютный прирост*** (**AAG;** *средняя скорость роста*) временного ряда вычисляется по следующей простой формуле:

$$\mathbf{AAG} = \frac{\sum_{k=1}^{n-1} \mathbf{AIC_k}}{\mathbf{n-1}} = \frac{\mathbf{L_n - L_1}}{\mathbf{n-1}} \qquad (68)$$

Данная формула – результат прямого применения метода средних к показателю *абсолютного прироста* **AIC**. Для примера табл. **17** получаем значение **AAG =** (95 - 5)/(28 - 1) = 3.21.

***Средний коэффициент роста*** (**ACG**) вычисляется как средняя геометрическая (*раздел* **6.2**) по следующей простой формуле:

$$\mathbf{ACG} = \sqrt[n-1]{\prod_{k=1}^{n-1} \mathbf{GC_k}} = \sqrt[n-1]{\mathbf{L_n/L_1}} \qquad (69)$$

Данная формула – результат применения метода средних к показателю *коэффициента роста* **AIC**. Для примера табл. **17** получаем значение $\mathbf{ACG} = \sqrt[27]{95/5} = \mathbf{1.12}$. Для вычисления корней высоких степеней используем прием логарифмирования с последующим потенциированием результата, а именно: $\mathbf{a} = \sqrt[n]{b} \Rightarrow \mathbf{ln(a) = b/n = c} \Rightarrow \mathbf{a = exp(c)}$. Однако, использование компьютеров сделало данную процедуру архаичной (*obsolete*).

***Средний темп роста*** (**AGR**) представляет собой величину **AGR=ACG*100%,** выражаемую в процентах. Однако, для практического применения показатель **AGR,** рассчитанный по данным о *конечном* и *начальном* уровнях ряда, можно использовать только в случае более или менее равномерного изменения уровней. В противном же случае использование средней геометрической может приводить к серьезным просчетам. ***Среднегодовой темп прироста*** (**AARG**) определяется на основе данных о *среднегодовых темпах роста*. Он вычисляется по формуле **AARG = AGR - 100%** и показывает, на сколько процентов в среднем изменялись уровни временного ряда. По данным примера табл. **17** (графы **1** и **2**) мы получаем величину **AARG =** (1.12 - 1)*100% = 12%. Полученная величина характеризует достаточно высокий среднегодовой темп прироста цитируемости в СССР научных публикаций **TTT**.

Рассмотренные показатели динамики имеют весьма широкое применение в статистических практике и исследованиях, их применение составляет основное содержание первых двух этапов анализа временных рядов. Показатели позволяют выявлять скорость и интенсивность развития явления, описываемого впеменным рядом. Дальнейший анализ ряда связан с более сложными обобщениями – определением основных компонент рядаЮ а именно: *трендовой*, *циклической*, *сезонной*, *разовой* и других. Данные вопросы рассматриваются в конце главы.

## 8.4. Выявление основной тенденции (тренда) временного ряда

*Временной* (*динамический*) ряд подвержен влиянию факторов *эволюционного* и *осцилляционного характеров*, а также различным *разовым воздействиям*. Под *эволюционным фактором* понимается ***тренд*** динамики, представляющий долго проявляющуюся основную тенденцию временного ряда. ***Осцилляционный*** фактор определяет сезонные, конъюнктурные и иные колебания ряда. ***Разовый*** фактор определяет спорадические воздействия на динамику, вызываемые, например, резким изменением среды развития явления. Таким образом, первоначальные значения временного ряда подвергаются самым разнообразным воздействиям и с учетом сказанного можно выделять четыре основные составляющие временного ряда : ***трендовую*** (**T**), ***циклическую*** или ***конъюнктурную*** (**C**), ***сезонную*** (**S**) и ***разовую*** (**O**). В общем же случае временные ряды характеризуются отмеченными четырьмя составляющими, а именно:

- ***тренд*** (**T**), *который представляет непериодическое изменение в среднем на временном интервале, на котором определен временной ряд*
- *один или более* ***сезонных*** *факторов* (**S**), *которые являются действиями, повторяющимися в единицах дней, недель, месяцев или лет*
- *другие* ***циклические*** *факторы* (**C**), *влияющее на анализируемое явление*
- ***случайный*** *фактор* (**O**), *являющийся результатом суммарного эффекта факторов, не идентифицированных в определении модели временного ряда*.

***Тренд, сезонные*** и ***циклические*** факторы – детерминированные составляющие модели ряда, в то время как ***случайная*** составляющая называется стохастической из-за непредсказуемости ее значений. Попытка распознать на глаз индивидуальные составляющие может быть довольно трудна. Между тем, существуют очень *широко* распространенные статистические методы для определения возможной структуры временного ряда, определенного изучаемым явлением.

Разложение временного ряда на составляющие позволяет представлять его в виде функции **M = Ψ(T, C, S, O)** от 4 переменных – основных факторов, влияющих на изучаемое явление. Принципиальная **M**-модель определяет способ, при котором сочетаются *детерминированные* компоненты (**T, C** и **S**) и *стохастическая* составляющая (**O**). Как правило, из-за сложности оценки влияния *случайного* фактора на *динамику* исследуемого явления, классические модели *временных* (*динамических*) рядов *не принимают его во внимание*. Однако, в *последние* годы в этом направлении наметился определенный прогресс, в значительной степени обусловленный *робастными методами* анализа временных рядов. В настоящее время, робастные методы составляют достаточно важный раздел современной математической статистики, с которой заинтересованный читатель может ознакомиться в превосходной книге [219], в свою очередь, содержащей обширную библиографию по данной проблематикех. Рассматриваемые ниже модели временных рядов не принимают во внимание случайный фактор; т.е. наши модели описываются посредством некоторой функции **M = Ψ(T, C, S)** от трех переменных. Имеется *два* традиционно используемых варианта для представления *зависимости факторов*, а именно:

**1**) **M = T + C + S** - ***аддитивная зависимость факторов*** (***модель***)

**2**) **M = T * C * S** - ***мультипликативная зависимость факторов*** (***модель***)

В зависимости от вида **Ψ**-функции можно говорить об *аддитивной* или *мультипликативной* **M-**модели временного ряда. В любом случае **M-**модель описывает разложение динамики изучаемогоо явления на его основные составляющие. *Аддитивная* **M-**модель определяется главным образом тем фактом, что его колеблющийся фактор имеет постоянный характер. В то время как для *мультипликативной* **M-**модели характер ее колеблющегося фактора остается постоянным только относительно *тренда* ряда. Существует много условий и предпосылок, определяющих выбор модели временного ряда. Создание хороших моделей существенно базируется на знании сути явления, описываемого временным рядом, а также на опыте по сглаживанию данных временными рядами. Тогда как общие комментарии предлагаются нами ниже, которые могут послужить полезным руководством и/или хорошим справочным материалом, чтобы избежать грубых ошибок при работе с временными рядами.

Различие между *аддитивными* и *мультипликативными моделями* весьма детально рассмотрено, например, в книгах [132, 285]. Общие соображения, лежащие в основе выбора того или иного типа **M-**модели, вообще говоря, сводятся к следующему. Потребность в *мультипликативной* модели ряда возникает тогда, когда величина *случайной изменчивости* растет с увеличением средней реакции. Как правило, это имеет место тогда, когда имеется значительная сезонная составляющая, которая завершается пиковой сезонной реакцией, *превышающей минимальные*

уровни ряда в десять или более раз. Более того, *мультипликативная* модель может быть также востребована при условии, что ряд имеет увеличивающийся *тренд* и проявляется *сезонная* и/или *циклическая* составляющая, увеличиваясь с увеличением значения тренда.

***Тренд*** – долговременная составляющая ряда, *основная тенденция* развития явления; при этом, все остальные составляющие рассматриваются как мешающие процедуре его определения. И коль скоро у нас имеется ряд наблюдаемых значений для различных моментов времени, то наша задача состоит в нахождении подходящей *трендовой* кривой, сглаживающей остальные колебания. Рассмотрим ряд методов нахождения трендовой кривой. Наиболее простым здесь является визуальный метод сглаживания, заключающийся в проведении через точки *уровней* временного ряда ломаной линии, отражающей основной характер тренда. Данный метод в любом случае имеет место при анализе ряда, позволяя сделать предварительные наметки для более точного сглаживания. В его реализации большую помощь может оказать *компьютерный* анализ, например, в среде пакета *Maple* или подобного ему программного средства.

В качестве наиболее простого можно предложить визуальный метод сглаживания, состоящий в проведении через точки (*уровни*) ряда ломаной линии, отражающей основной характер его тренда. При его реализации неоценимая помощь может быть предоставлена компьютерным анализом, например, в среде вышеупомянутого пакета *Maple* или подобного программного средства. Так, рис. **17** дает пример графического представления временного ряда с помощью пакета "*Diagram Microsoft Graph 97*" для *Windows*-платформы. Визуальный анализ *графического* представления ряда позволяет получать важную предварительную информацию (*основная тенденция*, *колебательный эффект*, *сезонность и т.д.*) о *временной динамике* изучаемого явления, отражаемого данным временным рядом.

При рассмотрении уровней за относительно короткие промежутки времени в силу влияния различного рода факторов наблюдается колебание, порой значительное, их значений, что затрудняет выявление тренда ряда. Поэтому само собой напрашивается ***метод укрупнения интервалов,*** состоящий в объединении нескольких смежных интервалов в один, например, вместо ежегодных берутся данные за пятилетия. При этом, часто применяется не общая величина уровня укрупненного интервала, а среднегодовой показатель, что еще лучше (*в силу специфики средних величин*, *см. разделы* **6.1, 6.4**) акцентирует общую тенденцию, сглаживая различного рода колебания динамики. Для иллюстрации материала данного раздела снова используем пример табл. **17** (графы **1** и **2**), сведя все необходимые расчетные вычисления в следующую разработочную табл. **18.**

**Таблица 18. Разработочная таблица для временного ряда, определенного табл. 17**

| Год | $L_k$ | $\bar{L}'_k$ | $\bar{L}''_k$ | $\bar{L}'''_k$ | $\tilde{L}_k$ | MA3 | MA5 | MA7 | MA9 |
|---|---|---|---|---|---|---|---|---|---|
| (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
| 1973 | 5 | | | | | - | - | - | - |
| 1974 | 7 | | | | | 8.0 | - | - | - |
| 1975 | 12 | | | | | 10.7 | 10.2 | - | - |
| 1976 | 13 | 11.4 | | | | 13.0 | 12.0 | 11.4 | - |
| 1977 | 14 | | | | | 13.7 | 13.6 | 13.3 | 13.9 |
| 1978 | 14 | | 17.7 | 17.7 | 18.62 | 14.3 | 14.8 | 16.1 | 16.0 |
| 1979 | 15 | | | | | 15.7 | 17.6 | 17.9 | 17.7 |
| 1980 | 18 | | | | | 20.0 | 19.6 | 19.1 | 19.6 |
| 1981 | 27 | | | | | 23.0 | 21.2 | 21.3 | 22.8 |
| 1982 | 24 | 24.0 | | | | 24.3 | 24.0 | 25.3 | 26.6 |
| 1983 | 22 | | | | | 25.0 | 28.8 | 30.0 | 31.2 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **1984** | **29** | | | | | **31.0** | **33.0** | **35.4** | **36.7** |
| **1985** | **42** | | | | | **39.7** | **39.4** | **40.7** | **42.2** |
| **1986** | **48** | | | | | **48.7** | **47.8** | **47.0** | **46.8** |
| **1987** | **56** | **55.6** | | | | **56.0** | **55.6** | **53.6** | **51.9** |
| **1988** | **64** | | | | | **62.7** | **60.8** | **59.4** | **56.7** |
| **1989** | **68** | | **59.4** | | | **66.7** | **65.2** | **62.7** | **59.9** |
| **1990** | **68** | | | | | **68.7** | **67.0** | **64.1** | **61.3** |
| **1991** | **70** | | | | | **67.7** | **65.8** | **64.0** | **62.7** |
| **1992** | **65** | **63.2** | | **65.7** | **65.14** | **64.3** | **63.2** | **63.4** | **63.4** |
| **1993** | **58** | | | | | **59.3** | **61.6** | **62.7** | **63.8** |
| **1994** | **55** | | | | | **57.7** | **60.2** | **62.6** | **64.6** |
| **1995** | **60** | | | | | **59.3** | **60.6** | **63.3** | **67.6** |
| **1996** | **63** | | | | | **63.3** | **64.0** | **67.6** | **-** |
| **1997** | **67** | **72.0** | **72.0** | | | **68.3** | **72.0** | **-** | **-** |
| **1998** | **75** | | | | | **79.0** | **-** | **-** | **-** |
| **1999** | **95** | | | | | **-** | **-** | **-** | **-** |

На *первом этапе* группируем интервалы (графа **1**) исходного ряда по количеству лет: **"7-5-5-5"** и вычисляем для каждой группы *среднегодовую* ($\overline{\mathbf{L}}'_{\mathbf{k}}$ ; графа 3) на основе данных графы **2.** Так как число лет не кратно пяти, то первую группу составили из 7 лет, динамика в пределах которой характеризуется *неубывающим* ростом значений уровней. При вычислении средней, она относится к *середине* укрупненного интервала. При выведении средних по укрупненным интервалам *отклонения* в уровнях ряда, обусловленные *случайными причинами*, нивелируются и более четко проявляется действие *основных* факторов изменения уровней – *общая тенденция* (***тренд***). Данные $\overline{\mathbf{L}}'_{\mathbf{k}}$ *средних* для укрупненных интервалов (табл. **18,** графа **3**) позволяют уже более четко представить тренд ряда, чем его линейная диаграмма (рис. **20**). Находя теперь средние величины средних ($\overline{\mathbf{L}}'_{\mathbf{k}}$), получаем дальнейшее усреднение уровней ряда ($\overline{\mathbf{L}}''_{\mathbf{k}}$). Наконец, вычисляя средние предыдущих средних ($\overline{\mathbf{L}}''_{\mathbf{k}}$), мы получаем *новые* средние уровней ряда ($\overline{\mathbf{L}}'''_{\mathbf{k}}$) и только два значения, позволяющие провести через них прямую линию **V(X)** – ориентировочный тренд ряда (***Trend_1***). Таким образом, графы **3, 4** и **5** табл. **18** представляют значения средних $\overline{\mathbf{L}}'_{\mathbf{k}}$, $\overline{\mathbf{L}}''_{\mathbf{k}}$ и $\overline{\mathbf{L}}'''_{\mathbf{k}}$ соответственно. Грубым способом укрупнения интервалов является деление ряда пополам и вычисление по обоим половинам средней ($\widetilde{\mathbf{L}}_{\mathbf{k}}$) (графа **6;** табл. **18**). Построенная по полученным двум точкам прямая ***Trend_2*** довольно близка к ранее построенной прямой линии **V(X) = *Trend_1*** (рис. **23**).

Описанные методы *выравнивания* временных рядов достаточно неточны и несколько лучшие результаты, в общем случае, дает ***метод скользящей средней,*** суть которого заключается в следующем. Для ряда формируем *скользящие укрупненные интервалы*, содержащие одинаковое число уровней. Каждый последующий интервал получаем, постепенно сдвигаясь от начала ряда на один уровень. По таким образом организованным интервалам определяем средне интервальные, относящиеся к их серединам. Поэтому, при *сглаживании* методом *скользящей средней* технически удобнее скользящий интервал составлять из *нечетного* количества *уровней* исходного ряда. Проиллюстрируем принцип расчета *скользящих средних* на примере табл. **18**.

В качестве укрупненного скользящего интервала выбираем трехлетний период и вычисляем *скользящую среднюю* (**MA3**), которая относится ко второму году интервала. Последовательно выбираются временные интервалы: 1973-75, 1974-1975, 1975-1977 и т.д. до исчерпания ряда (*по*

*этому принципу скольжения и получил название сам метод выравнивания рядов*). Величины **МАЗ** представлены в графе **7** (табл. **18**). В графах **8, 9** и **10** табл. **18** приведены значения *скользящей средней* соответственно для **5-** и **7-** и **9-**летних периодов *исходного* временного ряда. Исходный и сглаженный с помощью *скользящей средней* (табл. **18**) ряд представлены на рис. **25,** который получен в среде известного пакета **"***Diagram Microsoft Graph 97***"**.

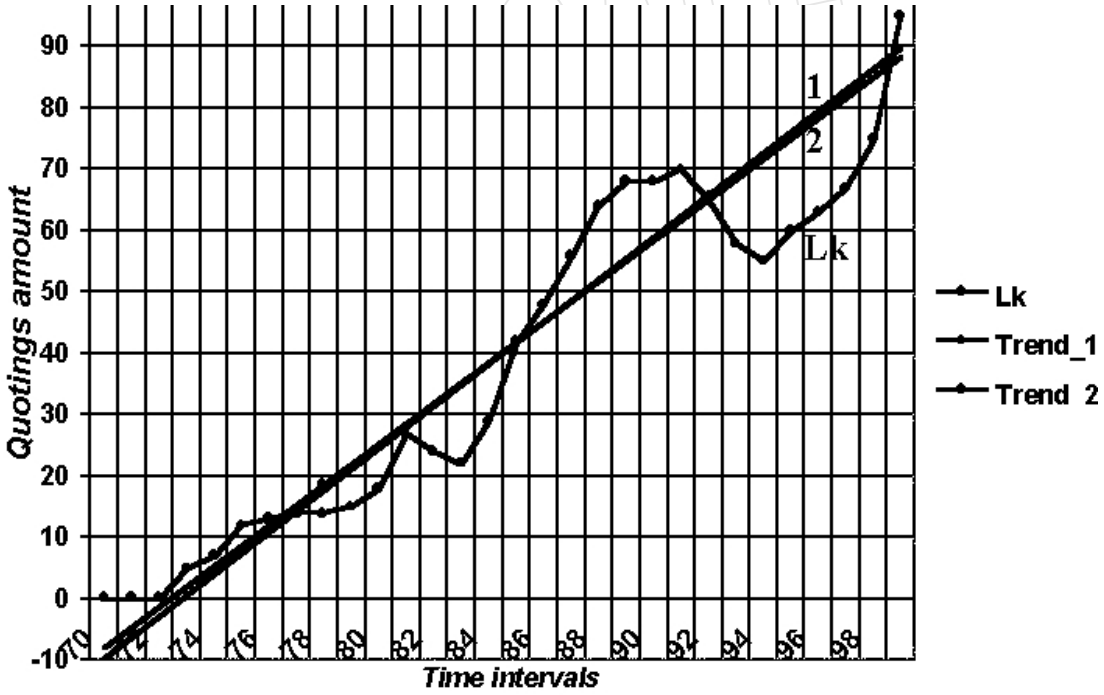



**Рис. 23. Приблизительное определение тренда ряда методом укрупнения интервалов**

Для автоматизации *вычисления скользящих средних* в среде *Maple* была реализована процедура **MAM(L, n)** с двумя формальными аргументами, где **L** – файл данных или список значений уровней исследуемого временного ряда и **n** – длина укрупненного скользящего интервала. Вызов процедуры возвращает список всех допустимых скользящих средних для отрезка скольжения длины **n** для заданного списка или текстового файла данных **L** статистических данных. При этом, если был закодирован третий дополнительный аргумент **p**, то результат возвращается с **p** значащими цифрами. Следующий фрагмент иллюстрирует применение процедуры для вычисления скользящих средних уровней ряда, определенного **Lk**-списком значений его уровней (рис. **24**). В качестве **Lк-**списка выбирается список значений уровней ряда, определенного табл. **18** (графа **2**), и как процедура **MAM** используется для вычисления *скользящих средних* относительно укрупненного скользящего интервала длины **9.** Полученные результаты соответствуют данным графы **10** табл. **18.**

```
MAM := proc(L::{list, file}, n::posint)
local a, b, k, p;
    `if`(type(L, list), assign(b = L), assign(b = evalf(convert(readdata1(L, 1), list))));
    assign(a = [ ]), [seq(assign('a' = [op(a), sum(b[p], p = k .. k + n - 1) / n]), k = 1 .. nops(b) - n + 1),
        `if`(nargs = 3 and type(args[3], posint) and 2 ≤ args[3], evalf(op(a), args[3]), op(a))]
end proc
> Lk:=[5,7,12,13,14,14,15,18,27,24,22,29,42,48,56,64,68,68,70,65,58,55,60,63,67,75,95]: MAM(Lk, 9, 4);
  [13.9, 16.0, 17.7, 19.6, 22.8, 26.6, 31.2, 36.7, 42.2, 46.8, 51.9, 56.7, 59.9, 61.3, 62.7, 63.4, 63.8, 64.6, 67.6]
```

**Рис. 24. Простая *Maple*-процедура для вычисления скользящих средних**

Данная процедура может быть полезна при решении статистических задач в среде *Maple*. Необходимо отметить, что эта и другие статистические *Maple*-процедуры, представленные в книге рассматриваются достаточно подробно в нашей недавней книге [302, 303].

Графическое представление *исходного* и *сглаженного* временного ряда (*посредством скользящих средних* **MA3, MA5, MA7** и **MA9;** табл. **18,** графы **1, 2 и 7 - 10**) может быть найдено на рис. **25.** Представление было получено посредством пакета "*Diagram Microsoft Graph 97*".

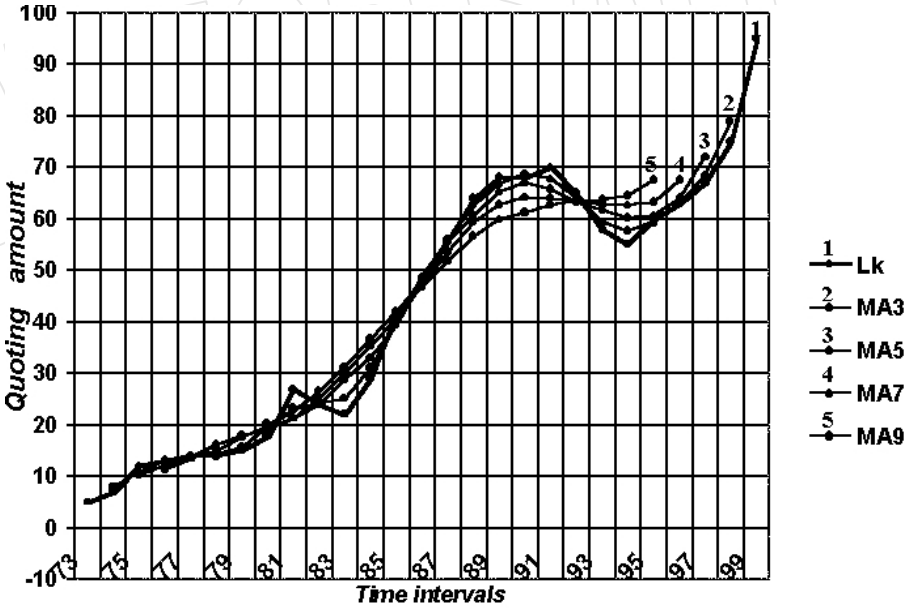



**Рис. 25. Графическое представление исходного и сглаженного временных рядов**

В данных, полученных методом *скользящих средних*, устраняется часть *колебаний* уровней ряда и их величины становятся более плавными относительно фактических уровней. Но данный метод имеет и свои недостатки, а именно: (**1**) невозможность получения уровней на концах сглаживаемого ряда (*за каждое сглаживание ряд уменьшается с обоих концов на величину* **(d - 1)/2,** *где* **d** *– длина скользящего интервала;* **d** *– четное*); (**2**) большая произвольность выбора интервала сглаживания, что может негативно влиять на выявление тренда; (**3**) наконец, в случае сильно осциллирующего характера динамики метод не дает сколько-нибудь хорошей картины за первую итерацию, что требует проведения ряда итераций.

Необходимо отметить, что вышеупомянутые методы *сглаживания* на сегодня имеют, главным образом, познавательный характер, так как массовое использование **ПК** решает подобные проблемы намного более эффективно. Однако, знание данных методов позволяет глубже проникать в сущность данной проблемы, тем более, что, например, метод *скользящей средней* реализован и программно ("*Diagram Microsoft Graph 97*" *и др*. [36, 42]).

Рассмотренные методы выравнивания (*сглаживания*) временного ряда дают ряд, отражающий некий *тренд* развития явления, более или менее свободный от случайных и осцилляционных колебаний. Для теоретической модели тренда ряда, лишенной указанных выше недостатков, используется ***аналитическое*** выравнивание. В этом случае уровни ряда выражïаются в виде аналитической функции от времени. *Аналитический* метод выравнивания является хорошей предпосылкой для применения других приемов углубленного изучения динамики явлений различного характера и природы.

При аналитическом выравнивании динамики ряда закономерно изменяющиеся его уровни представляются функционально в виде **Y(t)=F(t),** где **Y(t)** – уровни ряда как *функция времени*.

Тип же **F**-функции определяется характером динамики и для заключения о нем может быть в первом приближении использована линейная диаграмма временного ряда. В качестве **F**-функции выравнивания (*сглаживания*) может быть: *прямая линия* (*рост уровней ряда происходит в арифметической прогрессии*), *показательная кривая* (*когда динамика меняется в геометрической прогрессии*), *параболическая кривая*, *логарифмическая* и другие.

Основанием для выбора **F**-функции выравнивания должен служить содержательный анализ сущности *динамики* исследуемого явления. Во всяком случае за ним остается последнее слово. Из практических соображений для этих целей можно эффективно использовать *графическое* представление *экспериментального* ряда, что весьма удобно делать в рамках описанной выше компьютерной технологии. Примеры ее использования дают и рис. **17, 18, 19, 20, 22, 23, 25.** Компьютерный метод может включать визуализацию как линейной диаграммы ряда, так и результаты его сглаживания методами укрупненных интервалов, скользящих средних и др. Это позволяет в определенной мере нивелировать случайные и осцилляционные колебания временного ряда, затушевывающие его тренд.

При выборе типа кривой выравнивания ряда пользуются, как правило, методом ***конечных разностей,*** который основан на их свойствах относительно того или иного типа кривой. При этом, обязательным условием является равенство временных интервалов между уровнями временного. ***Конечными разностями*** от первого до (**n-1**)-го порядка называются величины **D(к)j,** определяемые следующими простыми рекуррентными формулами:

$$D(1)_j = L_{j+1} - L_j \;\ldots\; D(k)_j = D(k-1)_{j+1} - D(k-1)_j \qquad \{k = 1..n-1;\ j = 1..n-1\}$$

На основе данных формул легко выражать разности **D(к)j k**-го порядка [*в количестве* (**n - k**)] через разности первого порядка ($L_{j+1} - L_j$) и коэффициенты формального бинома $(1-1)^k$. Из свойств конечных разностей относительно того или иного вида кривой, в частности, следует, что если **F**-функция выравнивания временного ряда имеет вид:

**1**) **F(t) = A*t + B,** *то первые разности постоянны, тогда как вторые равны нулю*

**2**) **F(t) = A*t$^2$ + B*t + C,** *то вторые разности постоянны, тогда как третьи равны нулю.*

Для вычисления конечных разностей любого порядка уровней временного ряда мы можем использовать простую процедуру **FD(L),** которая на основе **L**-списка значений уровней ряда возвращает последовательность списков разностей всех возможных порядков для **L**-ряда, включая список уровней в качестве первого элемента. Исходный текст процедуры наряду с примером ее применения для ряда, определенного табл. **18,** представлен на рис. **26**.

```
> Lk:= [5, 7, 12, 13, 14, 14, 15, 18, 27, 24, 22, 29, 42, 48, 56, 64, 68, 68, 70, 65, 58, 55, 60, 63, 67, 75, 95]:
FD := proc(L::list)
local k, p, FD1, A, K;
    FD1 :=
        K → [assign('A' = [ ]), seq(assign('A' = [op(A), K[p + 1] - K[p]]), p = 1 .. nops(K) - 1), op(A)];
    L, FD1(L), seq(FD1(A), k = 2 .. nops(L) - 1)
end proc
> FD(Lk);
```

[5, 7, 12, 13, 14, 14, 15, 18, 27, 24, 22, 29, 42, 48, 56, 64, 68, 68, 70, 65, 58, 55, 60, 63, 67, 75, 95], [2, 5, 1, 1, 0, 1, 3, 9, -3, -2, 7, 13, 6, 8, 8, 4, 0, 2, -5, -7, -3, 5, 3, 4, 8, 20], [3, -4, 0, -1, 1, 2, 6, -12, 1, 9, 6, -7, 2, 0, -4, -4, 2, -7, -2, 4, 8, -2, 1, 4, 12], [-7, 4, -1, 2, 1, 4, -18, 13, 8, -3, -13, 9, -2, -4, 0, 6, -9, 5, 6, 4, -10, 3, 3, 8], [11, -5, 3, -1, 3, -22, 31, -5, -11, -10, 22, -11, -2, 4, 6, -15, 14, 1, -2, -14, 13, 0, 5], [-16, 8, -4, 4, -25, 53, -36, -6, 1, 32, -33, 9, 6, 2, -21, 29, -13, -3, -12, 27, -13, 5], [24, -12, 8, -29, 78, -89, 30, 7, 31, -65, 42, -3, -4, -23, 50, -42,

10, -9, 39, -40, 18], [-36, 20, -37, 107, -167, 119, -23, 24, -96, 107, -45, -1, -19, 73, -92, 52, -19, 48, -79, 58], [56, -57, 144, -274, 286, -142, 47, -120, 203, -152, 44, -18, 92, -165, 144, -71, 67, -127, 137], [-113, 201, -418, 560, -428, 189, -167, 323, -355, 196, -62, 110, -257, 309, -215, 138, -194, 264], [314, -619, 978, -988, 617, -356, 490, -678, 551, -258, 172, -367, 566, -524, 353, -332, 458], [-933, 1597, -1966, 1605, -973, 846, -1168, 1229, -809, 430, -539, 933, -1090, 877, -685, 790], [2530, -3563, 3571, -2578, 1819, -2014, 2397, -2038, 1239, -969, 1472, -2023, 1967, -1562, 1475], [-6093, 7134, -6149, 4397, -3833, 4411, -4435, 3277, -2208, 2441, -3495, 3990, -3529, 3037], [13227, -13283, 10546, -8230, 8244, -8846, 7712, -5485, 4649, -5936, 7485, -7519, 6566], [-26510, 23829, -18776, 16474, -17090, 16558, -13197, 10134, -10585, 13421, -15004, 14085], [50339, -42605, 35250, -33564, 33648, -29755, 23331, -20719, 24006, -28425, 29089], [-92944, 77855, -68814, 67212, -63403, 53086, -44050, 44725, -52431, 57514], [170799, -146669, 136026, -130615, 116489, -97136, 88775, -97156, 109945], [-317468, 282695, -266641, 247104, -213625, 185911, -185931, 207101], [600163, -549336, 513745, -460729, 399536, -371842, 393032], [-1149499, 1063081, -974474, 860265, -771378, 764874], [2212580, -2037555, 1834739, -1631643, 1536252], [-4250135, 3872294, -3466382, 3167895], [8122429, -7338676, 6634277], [-15461105, 13972953], [29434058]

**Рис. 26. Вычисление конечных разностей для уровней ряда, определенного табл. 18**

Используя описанный подход к выявлению типа **F**-функции выравнивания для примера табл. **18,** обнаруживаем нарастание (*по абсолютной величине*) значений конечных разностей с чередованием знака через одну разность, начиная с конечных разностей **8**-го порядка (рис. **26**). Величина единственной разности 26-го порядка составляет 29.434.058. Следовательно, следуя теории, ни *линейная*, ни *параболическая* **F**-функции в качестве *тренда* для нашего ряда (табл. **18**) не подходят.

Объясняется это влиянием на *динамику* явления – *цитируемость работ* **ТТГ** *в СССР* – довольно сильных факторов, серьезно усложняющих выявление его эволюционного (*трендового*) фактора (рис. **23, 25**). Поэтому принимается решение выявить для данного ряда линейный тренд **F(t) = A*t + B,** чтобы относительно его более рельефно отразить другие факторы, влияющие на динамику. Для определения коэффициентов **A** и **B F**-функции используется хорошо известный *метод наименьших квадратов* (**МНК;** *прежде всего, в форме линейной модели регрессии*), согласно которому **F**-прямая выбирается так, чтобы минимизировать сумму квадратов отклонений фактических уровней ряда от их значений, вычисленных согласно **F**-функции в дискретные моменты времени **t = {k}** (**k = 1 .. n**), а именно:

$$\sum_{k=1}^{n}(L_k - F(k))^2 = \sum_{k=1}^{n}(L_k - A*k - B)^2 = \textit{minimum} \qquad (70)$$

Используя классические средства анализа [111, 166, 220] для определения *минимакса* функции от нескольких переменных {**A, B**} и решая относительно них полученную систему из 2-х линейных уравнений, отражающих *необходимое* и *достаточное* условие минимума, получаем:

$$A = \frac{6*\sum_{k=1}^{n}(2*k - n - 1)*L_k}{n^3 - n} \qquad B = \frac{2*\sum_{k=1}^{n}(2*n - 3*k + 1)*L_k}{n*(n-1)} \qquad (71)$$

В качестве полезного упражнения читателю рекомендуется вывести соотношения (71).

Следует отметить, что **МНК** дает оценки с *наименьшей* возможной дисперсией (*эффективные оценки*), если ошибки наблюдения и независимы, и подчиняются нормальному закону распределения. В этом смысле **МНК** является наилучшим среди всех остальных методов, позволяющих получать *несмещенные* оценки. Однако, если распределение случайных ошибок существенно отличается от нормального, то **МНК** может и не быть оптимальным. Сделанное замечание следует иметь в виду при любом применении данного метода.

Для дальнейшей иллюстрации анализа временного ряда рассмотрим примеры трех рядов **U, A** и **G** цитируемости работ **ТТГ** по **МТОС** и ее приложениям соответственно в СССР, за рубежом и итоговой цитируемости. Так как цитируемость, естественно, отстает по времени от публикаций, то в качестве *базисного* нами взят 1971 г. – начало цитируемости за рубежом и все три временных ряда получили единую точку отсчета на временной шкале. Используя теперь соотношения (71) и значения для уровней $\mathbf{L_k}$ (**k = 1..29**) рядов **U, A** и **G** (табл. **9;** графы **4, 5** и **7** соответственно), в среде пакета *Maple* определяем их линейные тренды и изображаем их графически совместно с линейными диаграммами этих рядов (рис. **27**).

```
> restart:  LT:= proc(L::list, X::symbol, t)
local k, n;
   n := nops(L);  evalf(6*sum((2*k - n - 1)*L[k], k = 1 .. n)*X/(n^3 - n) + 2*sum((2*n - 3*k + 1)*L[k],
         k = 1 .. n)/(n^2 - n), t)
end proc:
> U:= [0,0,0,5,7,12,13,14,14,15,18,27,24,22,29,42,48,56,64,68,68,70,65,58,55,60,63,67,75,95]:  A:= [0,33,
  38,42,76,94,96,98,102,125,145,174,186,205,220,222,236,240,266,284,292,304,314,316,320,336, 345,384,
  395,405]:  G:=[0,33,38,47,83,106,109,112,116,140,163,201,210,227,249,264,284,296,330,352,360,374,
  379, 374,375,396,408,455,470,500]:  FU(x):=LT(U,x,3);  FA(x):=LT(A,x,3);  FG(x):=LT(G,x,3);
                FU(x) := 3.01*x - 8.15      FA(x) := 13.5*x + 0.818      FG(x) := 16.5*x - 7.55
> with(plots):  P1:=plot({FU(x), FA(x), FG(x)}, x=1..29, color=[blue, red, green]):  L1:=listplot(U,
  color = blue, linestyle=3, thickness=2):  L2:= listplot(A, color = green, linestyle=3, thickness=2):
  L3:=listplot(G, color=red, linestyle=3, thickness=2): T:=textplot({[28,485,`G `], [28,410, `A `],
  [28, 95, `U `]}):  display({P1, L1, L2, L3, T}, thickness = 3, axesfont = [TIMES, BOLD, 10],
  font = [TIMES, BOLD, 16], labels= [``, ``]);
```

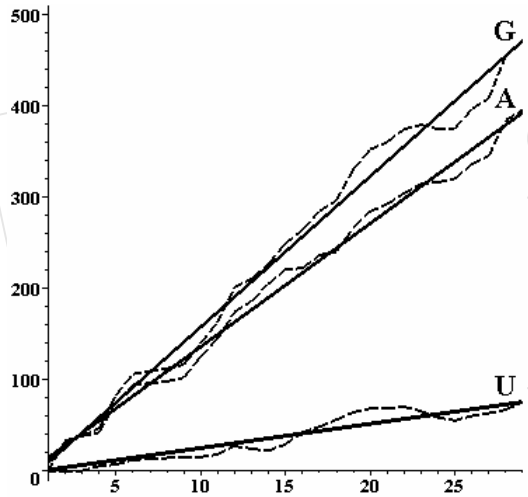



**Рис. 27. Вычисление линейных трендов рядов (табл. 9) с их графическим представлением**

Для вычисления *линейных трендов* вышеупомянутых временных рядов **U, A** и **G** в среде *Maple* реализована процедура **LT(L, X)** от двух формальных аргументов; ее первый аргумент **L** определяет значения уровней временного ряда, тогда как второй аргумент **X** – независимая переменная. Вызов процедуры возвращает уравнение для искомого линейного тренда **F (X) = A*X + B.** Приведенный *Maple*-фрагмент (рис. **27**) представляет и исходный текст процедуры, и результаты его применения для вычисления линейных трендов временных рядов **U, A** и **G**. В конце фрагмента приводятся графические представления самих временных рядов наряду с линейными трендами, соответствующими им.

Из полученных результатов *линейного выравнивания* рядов (*хорошо согласующихся с визуальным наблюдением*, рис. **27**) можно сделать ряд интересных выводов. Прежде всего, коэффициенты {3.01 *для* **FU(X)** *и* 13.5 *для* **FA(X)**} при **X**-переменной **F**-функции *выравнивания* (рис. **27**) говорят о существенном превышении абсолютного прироста (*так как в случае временного интервала ряда, равного* **1,** *величина коэффициента совпадает по значению с показателем абсолютного прироста*) цитируемости работ **ТТГ** за рубежом (**A**) относительно СССР (**U**). Во-вторых, цитируемость за рубежом идет с *опережением* как начала самого процесса в целом (*не взирая на то, что первые наши работы публиковались в трудах* АН Эстонии), так и его *колеблемости* относительно линейного тренда (рис. **27**).

Для сглаживания рядов посредством **МНК** используются самые различные типы **F**-функций (*линейные*, *параболы*, *гиперболы*, *показательные*, *степенные*, *логарифмические*, *тригонометрические* и др.). Для предварительной оценки типа искомой **F**-функции сглаживания эффективно использование описанной выше компьютерной технологии. На основе такого подхода в среде вышеупомянутых пакетов *MathCAD* [14, 15, 17], *Mathematica* [134, 135] и *Maple* [139-141, 143, 190] было выполнено *сглаживание* временного **U**-ряда на основе логарифмической **F**-функции $\mathbf{F(X) = \ln(X)/\ln(B)}$ с определением ее **B**-параметра на основе **МНК,** а именно:

$$\mathbf{B = \exp\left(\frac{\sum_{k=1}^{n} \ln^2(k)}{\sum_{k=1}^{n} \ln(k) * L_k}\right)}$$

Однако результат сглаживания оказался существенно хуже случая линейной **F**-функции. Не намного улучшило ситуацию и использование сглаживающей *степенной* **F**-функции. Однако, основной целью аналитического сглаживания ряда является не столько воспроизведение как можно точнее фактических данных, сколько определение модели развития исследуемого явления *во времени*. В этом плане вполне уместно ограничиться в наших примерах *линейными* трендами для **U-, A-** и **G-**рядов (рис. **27**).

***Интерполяция*** временного ряда состоит в вычислении недостающих внутренних уровней ряда и производится она распространением тенденции развития за известные временные интервалы на промежутки времени с отсутствующими данными. *Интерполяция* существенно упрощается при известном тренде ряда. В общем же случае интерполяция производится методами, хорошо известными из анализа [111, 188, 166, 220], включая графические методы на основе кривых выравнивания ряда. ***Экстраполяция*** состоит в вычислении уровней ряда вне наблюдаемого временного промежутка, но с использованием информации о нем на этом *промежутке*. Осуществляется она методами, подобными случаю *интерполяции*, и существенно облегчается при известном тренде ряда. Между тем, *экстраполяцию*, несмотря на кажущуюся очевидность, не следует рассматривать как завершающую стадию прогнозирования, а лишь как *предварительный* этап в разработке *прогноза* развития явления *во времени*. Для составления долговременного прогноза следует привлекать дополнительную информацию, отражающую динамику изучаемого явления и не содержащуюся в его ряде. Различают ***перспективную*** (*продолжение уровней ряда на будущее*) и ***ретроспективную*** (*продолжение уровней ряда в прошлое*) экстраполяции. Результаты интерполяции и экстраполяции временного ряда используются в регрессионном анализе и прогнозировании социально-экономических явлений различного характера [274, 276, 285].

Аналитическая **F**-функция тренда представляет собой математическую модель динамики явления и дает формульное выражение статистической закономерности, проявляющейся в ряде. Однако, следует иметь в виду, что метод аналитического выравнивания ряда содержит ряд условностей, связанных, в первую очередь, с тем, что уровни ряда рассматриваются как

**F(t)**-функция времени. Тогда как развитие явления обусловлено влиянием на него факторов, их направленности и интенсивности, а развитие явления во времени выступает как внешнее проявление этих факторов. Выявить *тренд* явления на основе **МНК** можно лишь тогда, когда выяснено, что изменяющиеся во времени процессы протекают на всем рассматриваемом временном промежутке одинаково, их *количественное* и *качественное* изменение происходит под влиянием одного и того же комплекса основных факторов.

Вместе с тем, в целом ряде случаев установить *справедливость* данного условия бывает весьма затруднительно. Примером могут служить рассматриваемые три ряда **U, A, G,** характеризующие цитируемость публикаций **ТТГ** по **МТОС,** которая определяется действием ряда сложно учитываемых факторов. ***Многофакторные динамические модели*** учитывают наиболее общие закономерности временной динамики явлений в изучаемый период времени и влияния на нее комплекса основных факторов. Вместе с тем, определение самого комплекса основных факторов в целом ряде случаев бывает весьма затруднительным.

## 8.5. Анализ случайной компоненты временного ряда

***Тренд*** определяет динамику временного ряда под влиянием основных факторов, однако фактические данные отклоняются от него в ту или иную сторону, образуя колеблющийся остаток **E(t).** Этот остаток называют ***случайной компонентой*** ояда и с его учетом уровень ряда представляется в виде *суммы систематической* (*ведущей*) *и случайной компонент*, а именно: **У(t) = F(t) + E(t),** где **F(t)** – тренд ряда. Анализ компоненты **E(t)** начинается с изучения ее колеблемости, определяемой колебанием фактических данных относительно *тренда*. Мерой колеблемости здесь выступает следующая величина:

$$Lt^2 = \frac{\sum_{k=1}^{n} (L_k - F(k))^2}{n} \qquad (72)$$

аналогичная дисперсии вариационного ряда (44), но в отличие от нее разности вычисляются не относительно средней, а относительно соответствующих значений *сглаживающей* **F**-кривой (*тренда*). Следует отметить, что именно минимизация этой величины лежит в основе **МНК**. Поскольку дисперсия выражается в квадратах единиц измерения уровней ряда, то мерой колеблемости служит показатель **Lt** – *среднее квадратическое отклонение*. В качестве примера вычислим показатели **Lt** для **A-** и **G-**рядов (рис. **28**).

```
> A:= [0,33,38,42,76,94,96,98,102,125,145,174,186,205,220,222,236,240,266,284,292,304,314,316,320,
  336, 345,384,395,405]:  G:= [0,33,38,47,83,106,109,112,116,140,163,201,210,227,249,264,284,296,330,
  352,360,374,379, 374,375,396,408,455,470,500]:  FA:= x -> 13.5*x + 0.818:  FG:= x -> 16.5*x - 7.55:
> Lt:= (L, F, t) -> evalf(sqrt(sum((L[k] - F(k))^2, k=1 .. nops(L))/nops(L)), t);
```

$$Lt := (L, F, t) \rightarrow \text{evalf}\left(\sqrt{\frac{\sum_{k=1}^{\text{nops}(L)} (L_k - F(k))^2}{\text{nops}(L)}}, t\right)$$

```
> Lt(A, FA, 6),  Lt(G, FG, 6);
```

10.8727,  15.5972

```
> Vt:= (L, F, t) -> evalf(Lt(L, F, t)/(sum(L[k], k=1 .. nops(L))/nops(L)), t);
```

$$Vt := (L, F, t) \rightarrow \text{evalf}\left(\frac{Lt(L, F, t)\ \text{nops}(L)}{\sum_{k=1}^{\text{nops}(L)} L_k}, t\right)$$

```
> Vt(A, FA, 4),  Vt(A, FA, 3)*100,  Vt(G, FG, 3),  Vt(G, FG, 3)*100;
                                0.05181,  5.1900,  0.0627,  6.2700
> H2:= (L, F, t) -> evalf(sum((L[k] - F(k))^2/F(k), k=1 .. nops(L)), t);
```

$$H2 := (L, F, t) \rightarrow \text{evalf}\left(\sum_{k=1}^{\text{nops}(L)} \frac{(L_k - F(k))^2}{F(k)}, t\right)$$

```
> I2:= (L, F, a, t) -> abs(evalf((H2(L, F, t) - nops(L))/sqrt(2*nops(L) + 4*a), t));
```

$$I2 := (L, F, a, t) \rightarrow \left|\text{evalf}\left(\frac{\text{H2}(L, F, t) - \text{nops}(L)}{\sqrt{2\,\text{nops}(L) + 4\,a}}, t\right)\right|$$

```
> Digits:=5:  I2(A, FA, 0, 3),  I2(G, FG, 0, 3);
                                0.685,  1.39
> R2:= (L, t) -> evalf(sum((L[k+1] - L[k])^2, k=1 .. nops(L) - 1)/(nops(L) - 1), t);
```

$$R2 := (L, t) \rightarrow \text{evalf}\left(\frac{\sum_{k=1}^{\text{nops}(L)-1} (L_{k+1} - L_k)^2}{\text{nops}(L) - 1}, t\right)$$

```
> R2(A, 5),  R2(G, 5),  sigma[At] = sqrt(R2(A, 5)/2),  sigma[Gt] = sqrt(R2(G, 5)/2);
                    301.97,  451.59,  σAt = 12.287,  σGt = 15.027
```

**Рис. 28. Вычисление *средне квадратичных отклонений*, *коэффициентов вариации* и некоторых других показателей для временных рядов A и G (табл. 9; *графы* 5 и 7 *соответственно*)**

Тренды для этих рядов были определены линейными **FA-** и **FG-**функциями соответственно (рис. **27**). Для вычисления **Lt**-показателя для обоих рядов создан соответствующий документ в среде *Maple* (рис. **28**), в котором определена простая процедура **Lt** наряду с некоторыми другими весьма полезными процедурами. В частности, вызов процедуры **Lt(L,F,t)** возвращает значение средне квадратичного отклонения для заданного ряда **L,** имеющего тренд **F,** с заданной **t**-точностью. Выполненное данного документа дает для временных рядов **A** и **G** следующие результаты: **Lt(A, FA, 6) =** 10.8727 и **Lt(G, FG, 6) =** 15.5972 соответственно.

Относительной мерой колеблемости является своеобразный ***коэффициент*** (**Vt**) ***вариации,*** вычисляемый по формуле **Vt=Lt/$\overline{L}$,** где $\overline{L}$ – средняя арифметическая уровней ряда. Для вычисления **Vt**-показателя для обоих рядов служит простая процедура **Vt**, чей вызов **Vt(L, F, t)** возвращает значение коэффициента вариации для заданного ряда **L,** имеющего тренд **F,** с заданной **t**-точностью. Выполнение *Maple*-документа (рис. **28**) дает для временных рядов **A** и **G** следующие результаты **Vt(A, FA, 5) =** 0.0518 (*или в процентах* 5.2%) и **Vt(Г, FG, 3) =** 0.0627 (*или в процентах* 6.3%) соответственно. С уменьшением значения **Vt**-показателя, функция тренда **F(t)** более точно отражает динамику ряда, поэтому данный показатель вполне может служить критерием адекватности выбора кривой тренда фактической динамике явления. В случае нашего примера выбор *линейного* типа тренда можно на первых порах считать вполне обоснованным.

Удостоверившись, в определенной мере, в приемлемости выбранного тренда ряда, следует проанализировать *степень случайности* **E(t)-**колебаний *ряда* относительно его *тренда*. В случае случайного характера колебаний величина **E(t)** должна иметь распределение, очень близкое к *нормальному*. Для проверки этой гипотезы воспользуется *критерием согласия* типа **χ-**квадрат

(см. раздел **2.5**). С этой целью вычисляем следующую величину **H2** χ-квадрата:

$$\mathbf{H2}=\sum_{k=1}^{n}\frac{(L_k-F(k))^2}{F(k)}$$

и используем критерий Ястремского, суть которого состоит в следующем. Вычисляется *величина* $\mathbf{I2=(H2-n)/\sqrt{2*n+4*a}}$ **,** где **n** – число групп (*элементов*) в *эмпирическом* распределении, а **a=a(t)** – величина, зависящая от числа групп, и при **n < 30** величина **a ≤ 0.6**. Полагая теперь величину **a = 0,** мы заранее завышаем значение для **I2**-показателя, что повышает степень достоверности выводов. Если **|I2| ≤ 3,** то эмпирическое распределение с вероятностью **P = 0.997** подчинено теоретическому **F**-распределению, в противном случае нет. Следовательно, при **|I2| ≤ 3** величина **E(t)** имеет нормальное распределение, т.е. носит случайный характер, иначе **E(t)-**отклонение вызвано отличными от случайных факторами. Выполненные в среде *Maple*-документа (рис. **28**) на основе процедур **H2** и **I2** (*сущность их формальных аргументов достаточно очевидны в контексте процедур* **Lt** и **Vt**) вычисления, для временных рядов **A** и **G** дают следующие значения для **I2**-показателя соответственно: **|IA2| = I2(A, FA, 0, 3) = 0.685 < 3** и **|IG2| = I2(G, FG, 0, 3) = 1.39 < 3,** говоря о случайном характере **E(t)**-изменчивостиь для рассматриваемых временных рядов.

Так как характер поведения величины **E(t)** имеет существенное значение для определения правильного выбора тренда временного ряда, то его анализ с целью получения некоторого предварительного заключения является дополнительным средством. С этой целью можно применять *показатель колеблемости* Ястремского, вычисляемый по следующей формуле:

$$\mathbf{R2}=\frac{\sum_{k=1}^{n-1}(L_{k+1}-L_k)^2}{n-1}$$

где **Lk** – уровни ряда (**k=1..n**). Показано, что между дисперсией $\sigma_t^2$ уровней и показателем **R2** существует соотношение: $\sigma_t^2=R2/2$**, i.e.** $\sigma_t=\pm\sqrt{R2/2}$ . Следовательно, до определения тренда мы получаем информацию по общей колеблемости уровней временного ряда. Вычисления, которые были выполнены в вышеупомянутом Maple-документе (рис. **28**) на основе простой процедуры **R2(L, t)** {*сущность ее формальных аргументов в контексте вышеупомянутых процедур* **Lt, Vt, H2, I2** *достаточно прозрачна*} для временных рядов **A** и **G** дают следующие значения *показателя изменчивости* $\sigma_{At}$ = 12.287 и $\sigma_{Gt}$ = 15.027 соответственно.

Располагая значением $\sigma_t$ до определения *тренда*, мы имеем возможность оценивать величину колеблемости компоненты **E(t).** Если она превосходит его общую колеблемость, то *тренд* ряда вычислен недостаточно точно и следует изменить тип сглаживающей кривой. Но это только общая рекомендация и решение следует принимать с учетом *специфики* изучаемого явления и его динамики. Так, для примера временных рядов **A** и **G** (рис. **28**) данная рекомендация не совсем выполняется, а именно: **Lt(A, FA, 6) <** $\sigma_{At}$ и **Lt(G, FG, 6) >** $\sigma_{Gt}$ (*относительно слегка*). Однако в виду *динамики* обоих рядов (рис. **20,23,27**) именно *линейные тренды* для них являются наиболее характерными, а колеблемость **E(t),** нося довольно случайный характер, вместе с тем, значительным образом объясняется и конъюнктурным фактором – периодическими подъемами и спадами активности в области исследований по **МТОС** и ее многочисленных приложений, что подтверждают и наши прежние логически-качественные анализы данной проблематики [4, 12, 23, 27, 191, 192, 274, 276, 285].

Таким образом, уровни временного ряда представляют собой сумму двух составляющих: ***тренда*** **F(t)** и ***случайного отклонения*** **E(t).** Первое слагаемое выражает наиболее характерные

черты явления, определяемого рядом, второе – случайные отклонения, вызванные действием множества различных по качеству, направленности и активности факторов (*разовые*, *сезонные*, *коньюнктурные* и др.).

## 8.6. Исследование периодических колебаний временного ряда

Целый ряд процессов носит *периодический* во времени характер. Прежде всего, это относится к различного рода сезонным явлениям, имеющим место во многих сферах человеческой деятельности. Типичным примером являются такие как: сельско-хозяйственные зработы, величина потребления элетроэнергии в течение суток и др. Под сезонностью понимают также неравномерность производственной деятельности в отраслях, связанных с сельским хозяйством. Сезонные колебания уровней ряда могут возникать в торговле из-за сезонного характера спроса на отдельные виды товаров, загрузке транспорта и др. Во многих случаях сезонность является весьма отрицательным фактором, например, из-за неравномерности использования оборудования и т.д. Статистическое исследование сезонности включает следующие основные задачи:

(**1**) *оценка количественного проявления сезонных колебаний*, *выявление их характера и интенсивности*

(**2**) *определение факторов*, *вызывающих сезонные колебания*

(**3**) *оценка последствий наличия сезонности*.

Для измерения сезонных колебаний имеется ряд методов, из которых наиболее простыми и употребительными являются такие методы: *абсолютных* и *относительных разностей*, *индексов сезонности*, *средних квадратических отклонений*, *аналитических моделей*. Мы кратко остановимся на ***аналитической модели,*** базирующейся на *гармонических* рядах Фурье [111,166, 220]. Данная модель позволяет описывать различного рода периодические процессы и ее аналитическое выражение применительно к временному ряду имеет следующий вид:

$$F(T) = A_0 + \sum_{k=1}^{m} (A_k * \cos(k*T) + B_k * \sin(k*T)) \qquad (73)$$

где **k** – номер гармоники ряда Фурье (*величина* **k** *чаще всего принимает значения от* 1 *до* 4), а параметры $\mathbf{A_0}$, $\mathbf{A_k}$ и $\mathbf{B_k}$ определяются на основе **МНК,** т.е. путем минимизации **Z**-функции:

$$Z = \sum_{k=1}^{n} (L_k - F(T_k))^2; \qquad A_0 = \frac{\sum_{k=1}^{n} L_k}{n} \qquad A_k = \frac{2*\sum_{k=1}^{n} L_k * \cos(k*T_k)}{n} \qquad B_k = \frac{2*\sum_{k=1}^{n} L_k * \sin(k*T_k)}{n}$$

Для изучения, например, сезонных колебаний на протяжении года следует полагать **n = 12** по числу месяцев в году, а в качестве временной шкалы ряда выбираются точки $\mathbf{T_k = (k-1)*\pi;}$ в этом случае ряд представляет собой упорядоченный набор пар вида $\mathbf{<L_k, T_k>}$ (**k = 1 .. n**) типа <*уровень*, *период года*>. Сопоставляя фактические и аналитические данные, можно как численно, так и визуально посредством компьютерной технологии оценивать степень адекватности отражения выбранной моделью эмпирических уровней исследуемого ряда.

Размах сезонных колебаний месячных данных принято измерять ***индексом сезонности*** – отношением средней из фактических месячных уровней ряда к средней из вычисленных по выбранной **F**-модели данных по тем же месяцам, а именно: **Isears = Lk/F(Tk)** (**k = 1 .. n**). Таким образом, величина **Isears** различна для каждого месяца и зависит от метода выравнивания (*скользящая средняя*, *аналитический метод* и др.). Показателем ***силы сезонности*** (**SF**) служит

среднее квадратическое отклонение индексов сезонности (*выраженное в* %) от 100%**,** а именно величина вида: $\mathbf{SF} = \sqrt{\sum_{k=1}^{12} (\mathbf{Isears, k} - \mathbf{100})^2 / 12}$ . При уменьшении величины **SF** можно говорить об уменьшении сезонности исследуемого явления, и наоборот. Изучение глубины сезонных колебаний можно проводить и путем определения отношений отклонения фактических уровней ряда от выровненных к модельным отклонениям, принимаемым за базисные.

Наряду с сезонными, ряды подвержены колебательному фактору *конъюнктурного* характера, имеющему самое широкое толкование: колеблемость научной активности в той или иной области, моды, вкусов и др. В качестве примера рассмотрим *колеблемость цитируемости работ* **ТТГ,** в определенной мере характеризующей изменения активности научных исследований по **МТОС** и ее приложениям в целом. Естественно, данный подход к оценке активности не лишен ряда недостатков (*сложность получения полной информации по цитируемости*, *разброс качества и проблематики самих цитируемых работ*, *их доступность и др.*). Между тем, в виду международного признания научной деятельности **ТТГ** [23, 191, 192] на основе анализа цитируемости ее работ можно, пожалуй, в определенной мере выявить основные тенденции колебательного процесса научной активности по **МТОС** и ее приложїениям в целом. Анализ проводим методом *компьютерной технологии* в среде математического пакета *Maple* (рис. **29**).

```
> G:= [0,33,38,47,83,106,109,112,116,140,163,201,210,227,249,264,284,296,330,352,360,374,379, 374,
  375,396,408,455,470,500]:  F:=[TIMES,BOLD,10]:  FG:= x -> 16.5*x - 7.55:  E:= k -> G[k] - FG(k):
> with(plots):  I81:= listplot([[81, -38], [81, 38]], color=gold):  I89:= listplot([[89, -38], [89, 38]],
   color=gold):  M:= textplot([[71.5,-30,`a`], [81.5,-30,`b`], [89.5,-30,`c`], [98.5,-30,`d`]], font= [F[1],
  F[2], 14], color=blue):  Tp:= plot(0,x=71..99, thickness=3, color=red):  Ep:=listplot([seq([70+k,
  E(k)], k=1..29)], style=line, thickness=2,  color=green, axesfont=F, axes=boxed):  TxT:=
  textplot([86.3, -3.5, `Trend:  FG(x) = 16.5*x - 7.55`], font=F, color=red):  display([Tp, Ep, TxT, I81,
  I89, M], labels= [`Annual intervals`, `E(x)`], labelfont = F);
```

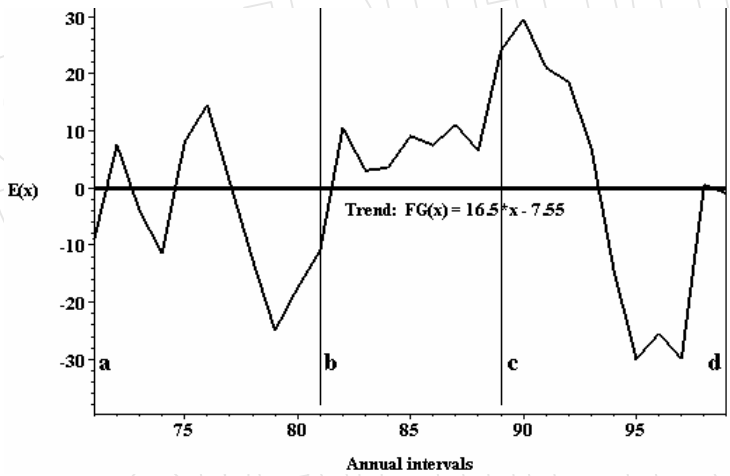



**Рис. 29. Колебательная составляющая временного ряда, определенного табл. 9 (графа 7)**

В качестве исследуемого выбирается **G**-ряд (табл. **9**, графа **7**), характеризующий динамику цитируемости работ **ТТГ** по **МТОС** в целом. Относительно полученного его тренда **FG(X)** (рис. **27**) определяется следующая функция отклонений: **E(X) = G(X) - FG(X),** колебательное поведение которой и анализируется. С учетом *размаха* колебаний **E(X)** выбираются масштабы осей координат и изображаются на одном графике тренд **G**-ряда (*прямая красного цвета* **– E(x) = 0**) и линейная диаграмма **E(X)**-функции отклонений (*ломаная зеленого цвета***;** рис. **29**). Из визуального анализа четко выделяются *три* основные *периода* колебания цитируемости работ

**ТТГ,** а именно: **<ab>** (1971-1981) – *синусоподобный*, **<bc>** (1982-1989) – *бессистемный* и **<cd>** (1988 - 1994) – **U**-*подобный*. При этом, определение периодов связано только с самим характером поведения на них **E(X)**-функции отклонений от трендовой линии.

Период **<ab>** характерен *периодическим* ростом и спадом активности исследований по **МТОС,** что нашло свое отражение в различных обзорах тех и последующих лет [1,4,23,127,191,192]. В качестве достаточно полезного упражнения читателю рекомендуется выровнять этот период гармоническим рядом Фурье (73). Период **<bc>**, в первую очередь, характеризуется ростом активности по прикладным аспектам **МТОС,** особенно в вычислительных, физических и биологических науках. Это подтверждает и *колеблемость* цитируемости выше уровня *тренда* ряда. Однако такой ее спорадический характер говорит и о многочисленных поисках новых областей приложения **МТОС** – нового перспективного раздела современной математической кибернетики [1, 4, 5, 25-28, 138, 190-192, 222]. Наконец, период **<cd>** характеризует новый этап резкого подъема активности по **МТОС** и ее уже достаточно многочисленным приложениям, интереса к данной проблематике со стороны многих теоретических и прикладных областей знаний, хотя его первая половина и говорит, на первый взгляд, об обратном. Но это связано, прежде всего, с тем, что постсоветские реалии последних лет не позволили нам проводить (*как прежде*) основательное статистическое наблюдение активности **ТТГ,** что сказалось и на использованных здесь данных за 1992-1994 годы деятельности группы.

Таким образом, анализ в значительной мере случайной функции **E(X)** колеблемости рядов, отражающих цитируемость научных работ, относительно своего *тренда* позволяет (*с учетом специфики и сути изучаемого процесса*) получать весьма интересную *информацию*, в частности, для генетической методологии науковедения – дисциплины, изучающей *функционирование* и *развитие науки*, *структуру* и *динамику научного* знания и *научной* активности, *взаимодействие* науки с другими сферами деятельности человека. Оформление *науковедения* в качестве *вполне* самостоятельной комплексной дисциплины можно отнести к 60-м годам прошлого столетия.

Как уже неоднократно отмечалось, модели *линейной регрессии* и *линейного тренда*, наряду с методом наименьших квадратов для их оценки получили самое большое распространение в статистическом анализе. Кроме того, классические оценки и процедуры проверки гипотез на случай линейных моделей также основаны на ***методе наименьших квадратов.*** Между тем, несмотря на математическую элегантность и непринужденность вычислений этого метода, он катастрофически страдает от недостатка ошибкоустойчивости, т. е. возможности учета резко выступающих наблюдений. Действительно, только одно резко выступающее значение может принести к непредсказуемым последствиям для оценки некоторого исследуемого явления. Поэтому, в книге [219] в связи с *ошибкоустойчивостью* исследован довольно *широкий* класс оценок *регрессии*, представляющих собой обобщение классического *метода наименьших квадратов*. Заинтересованный читатель может ознакомиться с этими и другими интересными проблемами ошибкоустойчивости в статистике в вышеупомянутой книге [219].

## 8.7. Сравнительный и связный анализы временных рядов

***Сравнительный анализ*** временных рядов состоит в сопоставлении во времени двух и более процессов ими описываемых. В параллельных сравнениях вновь возникают вопросы о сопоставимости уровней, но уже двух сравниваемых между собой рядов. Так как приведение к сопоставимому виду, например, по ценам и методике расчета показателей – дело весьма непростое, то статистика использует приемы сопоставления относительных, а не *абсолютных величин* (*глава* **5**). В этом случае данные о величине исследуемого показателя за год, принятый за *базисный*, берутся за 100%, а уровни остальных лет рядов соотносятся к нему. Примером приведения данных к одному основанию служит табл. **19,** в которой сгруппированы по

трехлеткам темпы цитируемости работ **ТТГ** по математической теории *однородных структур* и ее приложениям в различных областях современного естествознания относительно места самого источника цитирования, в качестве которого выбраны *две* большие группы, а именно: *отечественные* и *зарубежные* источники.

**Таблица 19. Трехлетние темпы цитируемости работ ТТГ по МТОС в СССР и за рубежом (1971 – 2000 г.г.; для 2000 г. *рассматривается только первое полугодие*)**

| *3-летние периоды цитирования* | *Отечественные цитирования* | | | *Зарубежные цитирования* | | |
|---|---|---|---|---|---|---|
| | **в % к 1971 - 1973** | *Кол-во цитирований* | *Кол-во цитирований* | **в % к 1971 - 1973** | *Кол-во цитирований* | *Кол-во цитирований* |
| **(1)** | **(2)** | **(3)** | **(4)** | **(5)** | **(6)** | **(7)** |
| **1971 - 1973** | **100** | **5** | **100.0** | **100.0** | **113** | **100.0** |
| **1974 - 1976** | **640** | **32** | **640.0** | **235.4** | **266** | **235.4** |
| **1977 - 1979** | **860** | **43** | **134.4** | **287.6** | **325** | **122.2** |
| **1980 - 1982** | **1380** | **69** | **160.5** | **446.9** | **505** | **155.4** |
| **1983 - 1985** | **1860** | **93** | **134.8** | **572.6** | **647** | **128.1** |
| **1986 - 1988** | **3360** | **168** | **180.6** | **656.6** | **742** | **114.7** |
| **1989 - 1991** | **4120** | **206** | **122.6** | **778.8** | **880** | **118.6** |
| **1992 - 1994** | **3560** | **178** | **86.4** | **840.7** | **950** | **108.0** |
| **1995 - 1997** | **3800** | **190** | **106.7** | **942.5** | **1065** | **112.1** |
| **1998 - 2000** | **4100** | **205** | **107.9** | **995.6** | **1125** | **105.6** |

В *относительных величинах*, выраженных в базисных темпах роста цитируемости работ **ТТГ** в СССР и за рубежом, *несопоставимость* уровней рядов в значительной степени нивелируется. Так, темпы роста по сравнению с базисным периодом (1971 - 1973) цитируемости в СССР (табл. **19;** графы **4** и **7**) значительно выше относительно зарубежных цитирований, однако их общее количество за рубежом существенно больше (графы **3** и **6**). Процент же прироста за текущее трехлетие по отношению к предыдущему имеет для обоих случаев тенденцию к сближению (графы **4** и **7**). Объясняется это тем фактом, что до начала деятельности **ТТГ** за рубежом велись довольно интенсивные исследования по этой проблематике, тогда как в СССР **ТТГ** явилась *первым* научным коллективом в данном направлении, именно с ее работ в значительной мере началась активизация *отечественных исследований* в данном направлении [1, 4, 5, 12, 120, 190-192, 220, 221]. Более того, нами введена основная российская терминология по теории однородных структур (*Cellular Automata*) [313, 314]. Параллельные сравнения дают более яркую картину, если рассматривать *среднегодовые темпы прироста* и *средние абсолютные приросты* цитируемости.Читателю рекомендуется проделать это на примере табл. **9** и **19** в качестве весьма полезного упражнения.

Под ***коэффициентом опережения*** (**LC**) понимают отношения темпов прироста, отношение абсолютных приростов показателей разных явлений за сравниваемые периоды времени. Правильнее, однако, данный коэффициент понимать как отношение общих темпов роста для сравниваемых рядов за исследуемый период времени. Так, для случая примера (табл. **19;** графы **3** и **6**) темп роста цитируемости работ **ТТГ** за период 1974-1997 г.г. (*за* 21 *год*) составил в СССР 190/5 = 38 раза, а за рубежом – только 1065/113 = 9.4 раза. Коэффициент опережения цитируемости в СССР по сравнению с зарубежным за этот период составил уже величину **LC =** 38/9.4 ≈ 4 раза. Однако, этот *коэффициент* следует *рассматривать* и *оценивать* в *совокупности* с другими показателями и факторами динамики сравниваемых процессов. Так, для примера табл. **19** темп роста цитируемости нагих работ за рубежом за указанный период составил 9.4 раза и это, на наг взгляд, предпочтительнее, чем 38 раз аналогичного показателя для СССР,

ибо активность исследований за рубежом по данной проблематике существенно выше как количественно, так и качественно.

При изучении развития процесса во времени часто возникает необходимость оценки степени зависимости изменений уровней двух или нескольких связанных друг с другом временных рядов различного содержания. Исследование взаимосвязи между явлениями, отражаемыми рядами, возможно только на основе логическо-качественного рассмотрения, доказывающего возможность наличия такой связи и дающего ей удовлетворительную интерпретацию. Тогда как *количественную* оценку связи между рядами можно получать средствами *корреляционного* анализа, рассмотренного в предыдущей главе книги.

Здесь мы лишь напомним понятие *коэффициента корреляции* (**CC**), который служит числовой характеристикой взаимосвязи совместного распределения двух случайных величин **X = {X1, X2, X3, ... ,Xn}** и **Y = {Y1, Y2, Y3, ... ,Yn}**, и вычисляется по следующей общей формуле (59):

$$\mathbf{CC(X,Y)} = \frac{\mathbf{M[X*Y] - M[X]*M[Y]}}{\sqrt{\mathbf{D[X]*D[Y]}}} \qquad (74)$$

где **M[X]** и **D[X]** – соответственно *математическое ожидание* и *дисперсия* случайной **X**-*величины*. Если (**X, Y**) – двумерная *нормально* распределенная величина, то |**CC**| измеряет *степень связи* **X** и **Y;** если же закон распределения отличается от *нормального*, то показатель |**CC**| измеряет *линейную* степень зависимости. Таким образом, величина |**CC**| **≤ 1** – мера линейной степени зависимости выходов **Xk** и **Yk** (**k = 1 .. n**) статистических наблюдений. Если **CC > 0** (**CC < 0**), то с *увеличением* значения одной из величин, вторая имеет величина тенденцию *увеличиваться* (*уменьшается*); чем ближе значение показателя |**CC**| к **1,** тем более сильная связь случайных переменных **X** и **Y** имеет место. Специальная раздел математической статистики (*робастая статистика*) достаточно подробно рассматривает проблемы устойчивости корреляционного анализа относительно резко резко выступающих значений наблюдения [219].

Формулы (59, 74) *корреляции* двух случайных величин и *интерпретация* результатов расчетов по ней достаточно просты, однако применение метода корреляции для связного анализа временных рядов имеет ряд особенностей, которые необходимо учитывать для получения достоверных оценок взаимосвязи между рядами. Прежде всего, в ряде фактором, влияющим на изменение уровней, является само *время*. Изменение *уровней* с течением времени приводит к явлению так называемой ***автокорреляции*** – влиянию текущего уровня на последующие. Поэтому корреляция между уровнями двух рядов правильно отражает степень взаимосвязи явлений, если в каждом из рядов отсутствует автокорреляция. Наиболее ярко *автокорреляция* (*если она имеется*) проявляется между близлежащими уровнями ряда, для чего вычисляется величина **CC** (59, 74) для двух последовательностей **Lf = {L2, L3, ..., Ln}** и **Lp = {L1, L2, ..., Ln-1}** уровней ряда, рассматриваемая как величина ***коэффициента автокорреляции*** (**ACC**).

Рассмотрим вычисление **ACC** для проверки наличия *автокорреляции* в двух временных рядах **U** и **A** цитируемости работ **ТТГ** соответственно в СССР и за рубежом (табл. **9;** графы **4** и **5**). Из логическо-качественного анализа следует, что на самом деле *уровни* этих *рядов* должны иметь вполне определенную связь, характеризуемую величиной **0 < CC ≤ 1.** Все вычисления будем проводить в среде математического пакета *Maple* (рис. **30**).

```
> Digits:=6: with(SimpleStat): U:= [0,0,5,7,12,13,14,14,15,18,27,24,22,29,42,48,56,64,68,68,70,65,58,
  55,60,63,67,75,95]: A:= [33,38,42,76,94,96,98,102,125,145,174,186,205,220,222,236,240,266,284,292,
  304,314,316,320,336,345,384,395,405]: ACC(U, 4), ACC(A, 4);
                                0.9780, 0.9960
```

**> LT(U, FU, t, 4), LT(A, FA, t, 4);**

TU(x) = 3.046*x - 5.894, TA(x) = 13.37*x + 16.38

**> EU:= t -> U[t] - 3.046*x + 5.894; EA:= t -> A[t] - 13.37*x - 16.38; CC(EU, EA), CC(U, A);**

EU:= t -> U[t] - 3.046*x + 5.894

EA:= t -> A[t] - 13.37*x - 16.38

0.377595, 0.958119

**> T:= plots[textplot]({[21.5, 12, `EA(t) `], [21, -15, `EU(t) `]}): P1:= plots[listplot](EA, color=green, thickness=2, linestyle=4): P2:=plots[listplot](EU, color=red, thickness=2): plots[display]([P1, P2, T], axesfont = [TIMES, BOLD, 11], font = [TIMES, BOLD, 14]);**

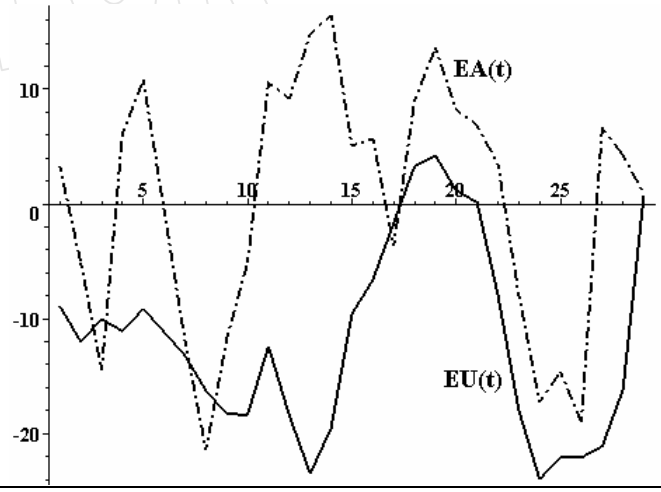



**Рис. 30. Вычисление *коэффициентов автокорреляции* для временных рядов U и A (табл. 9; графы 4 и 5) с выводом графиков осциллирующих остатков обоих рядов**

Для этого в среде *Maple* на основе формул (59, 74) реализована процедура **ACC.** Процедура экспортируется модулем **SimpleStat** наряду с другими средствами для проведения простого статистического анализа. Вызов процедуры **ACC(L)** возвращает значение **ACC** на основе **L-**списка уровней временного ряда. Посредством процедуры мы вычисляем значения **ACC** для вышеупомянутых рядов **U** и **A,** определенных данными табл. **9** (графы **4** и **5** соответственно). Коэффициенты автокорреляции для рядов **U** и **A** равны **ACC(U) =** 0.978 и **ACC(A) =** 0.996 соответственно (рис. **30**), т.е. для обоих связанных рядов автокорреляции являются очень большими. Следовательно, заключение о степени связи между уровнями этих рядов будет существенно искажено, если для его определения мы будем использовать непосредственно коэффициент корреляции. Для устранения влияния *автокорреляции* поступаем следующим образом.

В разделе **8.6** говорилось о том, что уровни **L(t)** временного ряда можно представить в виде **L(t) = F(t) + E(t),** где **F(t)** и **E(t)** – соответственно *тренд* и *колебательный остаток*, носящий, в основном, случайный характер. При этом, фактору *автокорреляции* подвержены именно **F(t)-**составляющие каждого ряда, тогда как величина **E(t)** ей не подвержена. Следовательно, для получения правильной картины связи между двумя временными рядами, не искаженной *автокорреляцией*, необходимо из уровней каждого ряда исключить их тренды и определить корреляцию **E(t)-**отклонений эмпирических рядов от их трендов.

Посредством процедуры **LT** мы получаем *линейные* тренды для рядов **U** и **A** соответственно следующего вида (рис. **30**):

**TU(t) = 3.046*x - 5.894, TA(t) = 13.37*x + 16.38**

Используем теперь их (*с учетом сказанного*) для получения функций **EU(t)** и **EA(t)** отклонений значений уровней соответственно рядов **U** и **A** от их трендов с последующим вычислением для них коэффициента корреляции (рис. **30**). Результаты такого вычисления дают несколько неожиданный результат – значение **CC(EU, EA)= 0.3776** говорит о достаточно слабой (*средней степени*) положительной корреляции обоих рядов, тогда как без учета *автокорреляции* налицо чрезвычайно тесная связь между рядами, а именно **CC(U, A)= 0.9581** (fig. **30**). Следовательно, без ее учета *автокорреляция* завысила бы показатель степени связи между рядами более, чем в **2.5** раза. Таким образом, на основе проведенного анализа нельзя говорить о наличии между временными рядами **U** и **A** достаточно тесной связи.

Вторая особенность применения метода корреляции к рядам состоит в возможном наличии для них ***временного лага*** – смещения во времени изменения одного явления относительно другого. ***Временной лаг*** (*лаг запаздывания*) – ***промежуток времени,*** за который изменение аргумента приводит к изменению результативного признака. Наличие *запаздывания* означает, что влияние переменной **X** на **Y-**переменную не проявляется немедленно, а растягивается на определенный промежуток времени. Примером такого лага может служить *запаздывание по времени* между научными результатами, их *цитируемостью* и *применением*. При этом, наряду с *постоянными*, рассматриваются лаги, *распределенные во времени*. Для исследования влияния такого типа лагов строятся их модели как линейные, так и нелинейные. Оценка параметров таких моделей осуществляется сведением их к ЛМР и широкому применению **МНК. МНК** является одним из наиболее распространенных методов обработки статистических данных, относящихся к различным функциональным зависимостям социально-экономических явлений. В том числе, он применим к **ЛМР** и позволяет получать достоверные оценки ее параметров, а также оценивать их погрешности. Кстати, многие прикладные компьютерные средства содержат метод наименьших квадратов.

Поэтому, при наличии *временного лага* в развитии явлений следует смещать уровни одного ряда относительно другого на *величину лага* и производить перерасчет *показателей корреляции*. Так, из *логических* рассмотрений развития **МТОС** в СССР относительно зарубежной научной активности налицо определенное *запаздывание* по времени, что позволяет говорить о наличии для рядов **U** и **A** лага в **4 - 5** лет, т.е. отставания на этот период активности исследований по **МТОС** в СССР. Сдвинув **U**-ряд назад относительно **A**-ряда на **4** года, произведем перерасчет *показателей корреляции* между ними (рис. **31**), т.е. относительно *сдвинутых* рядов производятся описанные выше расчеты: определение новых трендов, коэффициентов автокорреляции, вычисление величин **CC** с исключением автокорреляции и без оного. Все вычисления делаем в среде математического пакета *Maple* (рис. **31**).

```
> with(SimpleStat): Digits:= 6: A:= [33,38,42,76,94,96,98,102,125,145,174,186,205,220, 222,236,240,
  266,284,292,304,314,316,320,336]: U:= [12,13,14,14,15,18,27,24,22,29,42,48,56,64, 68,68,70,65,58,55,
  60,63,67,75,95]: CC(U,A), ACC(U), ACC(A);
                                0.933641, 0.969004, 0.995312
> n:= nops(A): LT(U, TU, x, 4), LT(A, TA, x, 4);
                        TU(x) = 3.105*x + 5.320, TA(x) = 13.44*x + 15.85
> FU:= x -> 3.105*x + 5.320: FA:= x -> 13.44*x + 15.85: EU:= [seq(U[t] - FU(t),t = 1 .. n)]:
  EA:= [seq(A[t] - FA(t), t = 1 .. n)]: CC(EU, EA);
                          3.105*t + 5.320, 13.44*t + 15.85, 0.229630
> T:= plots[textplot]({[23.7, 12, `EU(t) `],[22.4, -15, `EA(t)`]}): P1:= plots[listplot](EA, thickness=2,
  color = red, linestyle=4): P2:=plots[listplot](EU, color=blue, thickness = 2): plots[display]([P1,
  P2, T], axesfont = [TIMES, BOLD, 11], font = [TIMES, BOLD, 14]);
```

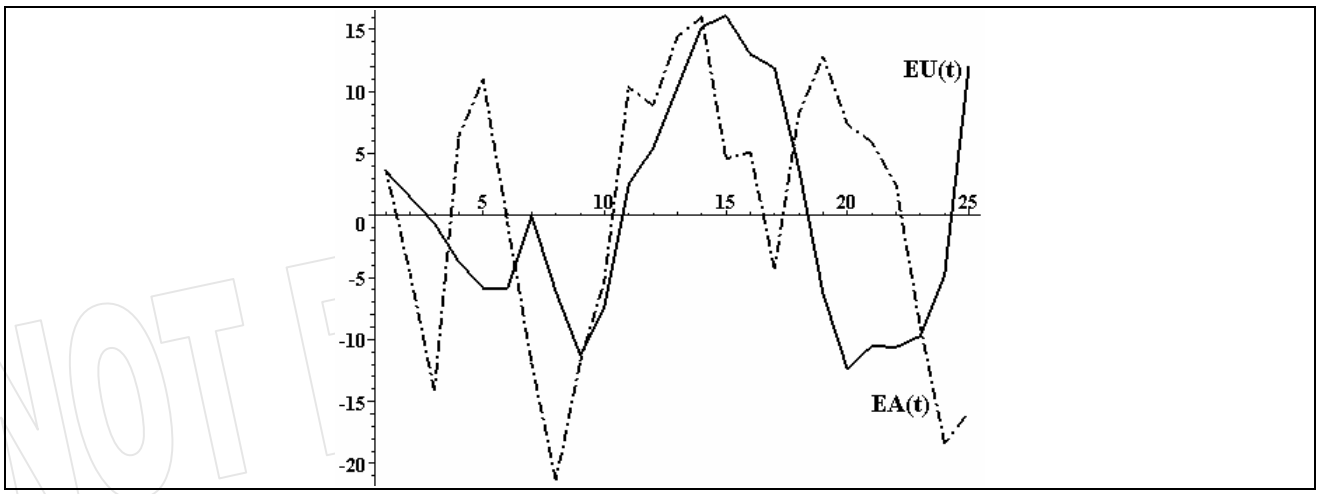



**Рис. 31. Вычисление для временных рядов A и U коэффициентов *корреляции*, *автокорреляции* и *линейных трендов* с учетом *временных лагов* наряду с вычислением CC-показателя после устранения *автокорреляции* и с выводом графиков колебательных остатков**

Результаты вычислений показывают, что для взаимно сдвинутых относительно выбранного *лага* (*в* **4** *года*) **U**-ряда коэффициенты автокорреляции для **U-** и **A-**рядов соответственно равны **ACC(U) =** 0.9690 и **ACC(A) =** 0.9953 (рис. **31**). При этом, относительно *аналогичных* показателей без учета лага эти значения незначительно уменьшились для обоих рядов (рис. **30**). Вместе с тем, *коэффициент корреляции* с исключенной из рядов *автокорреляцией* с учетом *лага* составил уже величину **CC(EU, EA) =** 0.2296, что даже несколько *ниже предыдущего* значения показателя {**CC(EU, EA) =** 0.3776; рис. **30**}. Тогда как коэффициент корреляции рядов с автокорреляцией теперь понизился до величины **CC =** 0.9336 (*предыдущее его значение было* 0.9581; рис. **30**), что также характеризует весьма тесную связь обоих рядов в целом. Таким образом, при весьма тесной связи динамики временных рядов **U** и **A** в целом, корреляционная связь между ними относительно невелика и положительна. Это может объясняться следующим фактором. При росте общего интереса к работам **ТТГ** по **МТОС** и ее приложениям взаимовлияние процессов цитируемости их в СССР и за рубежом не столь существенно. В принципе, данный результат имеет достаточно веские реальные подтверждения и обоснования.

Наконец, *третьей* особенностью *корреляции временных рядов* является возможность ***переменной корреляции*** **–** изменения коэффициентов *корреляции* со временем. Следовательно показатель степени связи рядов можно представлять как серию значений **CC,** вычисленных подобно методу *скользящей средней*. Наряду с этим, в сравниваемых рядах могут чередоваться периоды различной степени связи между ними, что объясняется действием целого ряда различных по силе, направленности и времени факторов. В качестве полезного упражнения читателю рекомендуется определить в рядах **U** и **A** (табл. **9;** графы **4** и **5**) периоды наибольшей связи между ними, для чего можно воспользоваться предложенной компьютерной технологией в среде известного математического пакета *Maple* либо другого подобного средства (*например*, *MathCAD*, *Mathematica*, *MatLab и др.*).

Исследование изменения связи между рядами во времени позволяет выяснять предпосылки, приводящие к изменению взаимосвязи между явлениями, отражаемыми этими временными рядами. Сказанное о корреляции в полной мере относится и к применению регрессионного метода для сравнения рядов. В любом случае, при оценке степени связи между временными рядами важнейшую роль играет логическо-качественный анализ связи между явлениями,

отражаемыми этими рядами, ибо в противном случае корреляция (*регрессия*) может показать не действительную зависимость, а чисто случайные, сопутствующие изменения динамики сравниваемых временных рядов.

Таким образом, при использовании методов *корреляции* и *регрессии* для временных рядов мы решаем двойную задачу измерения связи: (**1**) *последовательных уровней одного и того же ряда* и (**2**) *между соответствующими уровнями двух различных рядов, каким-то образом связанных друг с другом*. В *первом* случае вычисляются коэффициенты *автокорреляции* и *авторегрессии*, тогда как во втором – корреляции и регрессии. В первом случае коэффициенты вычисляются по непосредственным данным временных рядов, во втором – по их *отклонениям* от теоретически вычисленных трендов, *линейных* или *нелинейных*.

Временные ряды исследуются с различными целями. В одних случаях бывает достаточно получить описание характерных особенностей ряда, а в других случаях требуется не только предсказывать будущие значения временного ряда, но и управлять его поведением. Метод анализа временного ряда определяется, с одной стороны, целями анализа ряда, а с другой стороны, вероятностной природой формирования значений его уровней. Отметим только самые распространенные методы анализа временных рядов, а именно:

***Корреляционный анализ –*** позволяет находить существенные периодические зависимости и соответствующие им задержки (*лаги*) как внутри одного ряда (*автокорреляция*), так и между несколькими рядами (*кросскорреляция*).

***Спектральный анализ –*** позволяет находить периодические составляющие временного ряда.

***Модели авторегрессии и скользящего среднего*** ориентированы на описание процессов, проявляющих *однородные* колебания, возбуждаемые случайными воздействиями. Позволяют предсказывать будущие значения уровней временного ряда.

***Многоканальные модели авторегрессии и скользящего среднего*** применяются в тех случаях, когда имеется несколько коррелированных между собой временных рядов. В них имеются колебания, возбуждаемые одной причиной. Позволяют предсказывать будущие значения уровней временного ряда.

***Сезонная модель Бокса-Дженкинса*** применяется тогда, когда временной ряд содержит явно выраженный линейный тренд и сезонные составляющие. Позволяет предсказывать будущие значения уровней ряда. Данная модель была предложена в связи с анализом авиаперевозок.

***Прогноз экспоненциально взвешенным скользящим средним*** является простейшей моделью прогнозирования временного ряда. Применима во многих случаях, в том числе, охватывает модель ценообразования на основе случайных блужданий.

При этом, подробное рассмотрение упомянутых и других методов не представляется нам целесообразным, ибо как показывает практика, в простых (*модельных*) случаях весьма неплохо работают базовые алгоритмы, а в реальных ситуациях необходим индивидуальный подход и их соответствующие модификации. Более того, как следствие из опыта работы с временными рядами самой разной природы вытекает тот вывод, что практически в каждом случае можно подобрать метод анализа, дающий разумные результаты в рамках конкретной задачи. Но не всегда его можно строго обосновать или дать рекомендации по *выбору* метода. Следовательно, в достаточно сложных случаях разработка подходов к анализу данных формализуется плохо.

Заинтересованный читатель с вопросами анализа временных рядов и с современным состоянием в данном направлении может более детально ознакомиться, например, в следующих книгах [45, 126, 132, 157, 160, 172, 181, 182, 188, 219, 247-250, 257, 262, 273, 274, 276, 285, 315, 352].

# Глава 9.
# Элементы индексного метода анализа

В бизнесе, экономике и в других областях ***индекс*** – обычно не что иное, как отношение или среднее нескольких отношений; он используется, как правило, чтобы делать сравнения через какое-то время. По существу, рассматриваются два периода времени; один из них называют *базовым* периодом, а другой – *текущим* периодом. Число, относящееся к базовому периоду, служит основанием сравнения и появляется в знаменателе отношения (Ic/Ib), тогда как число, относящееся к текущему периоду появляется в числителе отношения (Ic/Ib). Важно помнить, что *индексы* – статистические оценки, построенные из данных, которые (*чаще да*, *чем нет*) являются элементами статистических выборок.

***Индексный метод*** представляет собой совокупность приемов, исторически возникших для измерения динамики экономических явлений: движения объемов продукции, динамики цен, производительности труда, себестоимости продукции и т. д. Индексные показатели выражаются относительными числами, а индексный метод в простейшем своем виде стал использоваться более 100 лет тому назад. Но серьезное развитие он получил значительно позднее, когда появились большие теоретические работы и практические исследования в этом направлении. Индексы стали широко применяться не только в анализе динамики явлений, но и для сравнения изменения явлений в пространстве и времени. *Индексный метод* в статистике – метод статистического исследования, основанный на построении и анализе индексов, позволяющих соизмерять сложные социально-экономические явления, особенно состоящие из непосредственно *несопоставимых* единиц. Специфика *индексного метода* состоит в том, что в статистическом ***показателе-индексе*** количественно несравнимые величины приводятся к некоторому общему единству, делающему их сравнимыми, соизмеримыми. Большую роль индексный метод играет в международной статистике при проведении различного рода международных сопоставлений.

## 9.1. Понятие индексов, их типы и назначение

***Индексом*** называется *относительный* статистический показатель, характеризующий *изменение* явлений во времени (***динамический индекс***) или в *пространстве* (***территориальный индекс***), например, *индексы* цен отдельных товаров, объемов различной продукции, ее себестоимости и т.п. В настоящее время статистика имеет дело с большим количеством различного рода и типа индексов, характеризующих те или иные стороны разнообразных явлений социально-экономического характера [68, 83, 86, 92, 96, 112, 113, 115, 125, 130, 157, 160, 173, 188]. Способы построения *индексов* зависят от *содержания* изучаемых явлений, *методологии* расчета исходных статистических показателей и целей самого исследования. Для удобства в теории статистики разработана специальная символика, в которой каждая измеряемая (*индексируемая*) величина имеет некоторое *символическое* обозначение, например: **q** – количество единиц данного вида продукции, **p, z** и t – соответственно цена, себестоимость и трудоемкость *единицы* продукции и др. Данная символика не является строго закрепленной и в ряде случаев может отличаться от приведенной выше, что может объясняться соображениями удобства.

По степени охвата элементов в совокупности индексы делятся на *индивидуальные* и *общие* (*сводные*, *сложные*). ***Индивидуальные*** *индексы* характеризуют изменение только *одного* элемента совокупности, например, изменение цитируемости научных работ только теоретического характера, изменение выпуска продукции определенного типа и др. Тогда как ***общий*** *индекс* характеризует изменение по всей совокупности элементов сложного явления, например, изменение цитируемости всех работ **ТТГ** в целом. Если индексы охватывают не все элементы *сложного* явления, а лишь часть его, то их называют ***групповыми*** или ***субиндексами,*** *например*, индексы продукции по отдельным отраслям промышленности. Следует отметить, что в рамках данной книги мы будем рассматривать ***общий*** индекс, как синоним индексов *сводного* и *сложного*. Тогда как в общем случае понятия *общего* и *сводного* индексов не совпадают, но для принятого нами *общего* уровня изложения данного раздела статистики такое отождествление вполне допустимо.

При вычислении индексов различают ***сравниваемый*** (***текущий***) уровень и ***базисный*** уровень, с которым производится сравнение. Выбор базиса определяется целью исследования и сутью изучаемого явления. В *индексации*, характеризующей *временные изменения* (*динамику*) явления, в качестве базиса используется показатель явления в каком-либо периоде, предшествующем текущему. При этом, возможны два способа *вычисления* индексов – *цепной* и *базисный*. ***Цепные*** индексы получаются сопоставлением текущих уровней явления с предшествующим, тогда как ***базисные*** – сопоставлением с уровнем некоторого фиксированного базисного периода. Для ***территориальных*** индексов за базис принимаются данные по какой-либо одной части территории или итоговый показатель по всей изучаемой территории в целом. При этом, при использовании индексов выполнения плана за базис берутся *плановые* показатели.

В зависимости от методологии расчета отличают ***агрегатные*** и ***средние*** от ***индивидуальных*** индексов. Последние, в свою очередь, подразделяются на *средние арифметические* и *средние гармонические* индексы. *Агрегатные* индексы качественных показателей могут вычисляться как индексы *переменного* и *постоянного* состава. В первом случае сопоставляются показатели, рассчитанные на базе изменяющихся структур явлений, во втором – на базе их постоянной структуры.

***Индексный метод*** играет довольно важную роль в социально-экономических исследованиях, характеризуя изменения уровней сложных общественных явлений. Так, *агрегатные* индексы можно использовать и в аналитических целях для оценки общего влияния на объемный показатель изменения образующих его факторов. Основной предпосылкой для проведения анализа *индексным методом* является возможность представления *суммарного* экономического показателя *произведением* двух или более определяющих его величину *показателей* (*факторов*) или *суммой* таких произведений. Например, объем выпуска продукции можно представить произведением уровня выработки одного работающего на средне списочную численность работников. Поэтому при анализе динамики выпуска продукции необходимо исследовать, в какой мере изменение объема выпуска продукции *вызвано* изменением каждого из указанных факторов.

Для оценки роли отдельных факторов изменения явления статистика широко использует ***системы взаимосвязанных индексов***. Здесь задача состоит в том, чтобы рассчитать изменение сложного показателя при изменении величины только одного фактора так, чтобы величина других факторов была бы сохранена на определенном постоянном уровне. В основе приемов аналитических индексных расчетов лежит принцип *абстрагирования* от изменений величины всех факторов, кроме исследуемого. При построении *индексов* или *их системы*, оценивающих влияние отдельных факторов на изменение сложного явления, необходимо иметь в виду, что общий результат изменения этого явления представляет собой суммарное изменение за счет влияния всех исследуемых, формирующих явление, факторов.

В ***индексной методологии*** недопустим формальный математический подход, что требует рассмотрения индекса прежде всего как экономического показателя. Задача индексного метода состоит в правильном измерении различных социально-экономических процессов. Выбор же математической формы того или иного индекса, его весов определяется *характером* исследуемых процессов. В условиях *социалистической* экономики *индексный метод* был одним из основных при решении задач планирования в народном хозяйстве. Многие плановые задания и анализ их выполнения использовали индексный подход. Весьма богатый опыт использования индексного метода в экономическом анализе накоплен в отечественных *отраслевых* статистиках, в первую очередь, в статистике промышленности. Следует отметить, что советская школа статистики требовала (*и не без основания*) увязки индексов с анализом описываемых ими социально-экономических процессов, тогда как зарубежная стоит, в основном, на формально-математических позициях, восходящих еще к работам И.Фишера по индексации. Хороший содержательный анализ обоих подходов к индексному методу можно найти в интересных работах [64-67,92,96,114-116,209]. Во многих отношениях советские индексные метод и методология позволяют отражать сравнительные аспекты изучаемых экономичных явлений с намного большей степенью адекватности [29, 64, 72, 82, 127, 128].

## 9.2. Индивидуальные и агрегатные индексы

***Индивидуальный индекс*** (**IIdx**) определяется путем *деления* величины **V1** некоторого элемента сложного явления в *текущем* периоде на его величину **Vo** в некотором *базисном* периоде, т.е. **IIdx(V)=V1/Vo.** Примеры вычислительных формул некоторых важных **IIdx** приведены ниже. Обязательным условием для вычисления **IIdx** является максимальная однородность явления, для которого он предназначен. Практически это условие выполнить очень трудно. Так, при определении индекса цен всегда уточняется к чему он относится, ибо в зависимости от вида продукции весьма велик разброс ее цен. **IIdx** является *относительной* величиной динамики и можїет быть *цепным* или *базисным*; выступают они в виде коэффициентов, процентов или промиле. Так, *индекс цитируемости* вида **IIQ = Q1/Qo** определяет изменение цитируемости работ в *текущем* периоде относительно *базисного*. Однако, **IIdx** совершенно непригодны для характеристики разнородных совокупностей в силу их специфики.

Применение ***общих индексов*** является дальнейшим развитием метода *средних* и обусловлено необходимостью характеризовать совокупности разнородных по содержанию элементов. Исходными величинами для построения *общих индексов* могут служить **IIdx**, *размеры* явлений, их специальные расчетные показатели. ***Агрегатный индекс*** (**AIdx**) – основная форма общих индексов и он характеризует относительные изменения индексируемой величины в *текущем* периоде относительно *базисного*. В общем виде **AIdx** *индексируемой* величины **X** вычисляются по следующим основным формулам:

$$\mathbf{AIdx(X)} = \frac{\sum_{k=1}^{n} \mathbf{Xc_k\ Wc_k}}{\sum_{k=1}^{n} \mathbf{Xb_k\ Wc_k}} \qquad (\mathbf{1})\ \textit{индекс Пааше}$$

$$\mathbf{AIdx(X)} = \frac{\sum_{k=1}^{n} \mathbf{Xc_k\ Wc_k}}{\sum_{k=1}^{n} \mathbf{Xb_k\ Wb_k}} \qquad (\mathbf{2})\ \textit{индекс Пааше} \qquad (75)$$

$$\mathbf{AIdx(X)} = \frac{\sum_{k=1}^{n} \mathbf{Xc_k\ Wb_k}}{\sum_{k=1}^{n} \mathbf{Xb_k\ Wb_k}} \qquad \textit{индекс Ласпейреса}$$

где **X** – индексируемая величина, **W** – веса индексов, **b** и **c** – знаки соответственно *базисного* и *текущего* периодов.

В качестве примера **AIdx** можно привести индекс *общей стоимости товаров* (*товарооборота*):

$$\mathbf{AIdx(P)} = \frac{\sum_{k=1}^{n} \mathbf{Pc}_k \mathbf{Qc}_k}{\sum_{k=1}^{n} \mathbf{Pb}_k \mathbf{Qc}_k} \tag{76}$$

где **P** и **Q** – соответственно цены и объемы товаров. Данный показатель является функцией изменения цен и отражает экономический эффект от изменения цен. Формула (76) дает **AIdx** *цен* с *текущими весами* и относится к типу *Пааше* (75.1). *Агрегатным* индекс называется потому, что берется определенная совокупность различных товаров, а с *текущими* весами потому, что *цены* взвешиваются по количеству текущей продукции. Отметим, в отечественных экономике и статистике используется именно этот тип **AIdx** (*индекс Пааше*), тогда как в зарубежной, как правило, употребляется индекс Ласпейреса (*с базисными весами*), т.е. в знаменателе формулы (76) вместо величины **Qc** (*текущей*) *используется* величина **Qb** (*базисная*). Естественно поэтому, различные конструкции индексов обусловливают и совершенно различные результаты относительно характера изменения как цен, так и объема, например, товорооборота.

Для иллюстрации вышесказанного сравним результаты вычисления по трем типам **AIdx** (75). Пусть нам известны объемы продаж товаров вида **G, S, Art** и **Kr,** и цены на них в 1995 и 2000 годах соответственно(см. табл. **20**). Требуется подсчитать **AIdx** цен указанных товаров.

**Таблица 20. Объемы продаж и цены**

| | *Товар* **G** | | *Товар* **S** | | *Товар* **Art** | | *Товар* **Kr** | |
|---|---|---|---|---|---|---|---|---|
| | **P** | **Q** | **P** | **Q** | **P** | **Q** | **P** | **Q** |
| *Базисный* **1995** | **48** | **947** | **28** | **967** | **6** | **989** | **2** | **996** |
| *Текущий* **2000** | **53** | **942** | **33** | **995** | **10** | **998** | **4** | **2005** |

Для этих целей в среде вышеупомянутого пакета *Maple* на основе формул (75), определяются функции **AI_P1(Xc, Xb, Wc, t), AI_P2(Xc, Xb, Wc, Wb, t)** и **AI_L(Xc, Xb, Wb, t)**, возвращающие значения индексов типов *Пааше* и *Ласпейреса* соответственно с заданной **t**-точностью (рис. **32**).

```
> Fract:= () -> evalf(sum(args[1][k]*args[3][k], k=1 .. nops(args[1]))/sum(args[2][k]*args[nargs - 1]
  [k], k = 1 .. nops(args[2])), args[-1]);
```

$$Fract := (\,) \rightarrow \mathbf{evalf}\left(\frac{\sum_{k=1}^{\mathbf{nops}(args_1)} args_{1_k}\, args_{3_k}}{\sum_{k=1}^{\mathbf{nops}(args_2)} args_{2_k}\, args_{nargs-1_k}}, args_{-1}\right)$$

```
> AI_P1:= (Xc, Xb, Wc, t) -> Fract(Xc, Xb, Wc, t):
> AI_P2:= (Xc, Xb, Wc, Wb, t) -> Fract(Xc, Xb, Wc, Wb, t):
> AI_L:= (Xc, Xb, Wb, t) -> Fract(Xc, Xb, Wb, t):
> AI_F:= (Xc, Xb, Wb, Wc, t) -> evalf(sqrt(AI_L(Xc, Xb, Wb, t)*AI_P1(Xc, Xb, Wc, t)), args[-1]);
```

$$AI\_F := (Xc, Xb, Wb, Wc, t) \rightarrow \mathbf{evalf}(\sqrt{\mathbf{AI\_L}(Xc, Xb, Wb, t)\,\mathbf{AI\_P1}(Xc, Xb, Wc, t)}, args_{-1})$$

```
> Pb:= [48, 28, 6, 2]:  Pc:= [53, 33, 10, 4]:  Qb:= [947, 967, 989, 996]:  Qc:= [942, 995, 998, 2005]:
  AI_P1(Pc, Pb, Qc, 4),  AI_P2(Pc, Pb, Qc, Qb, 4),  AI_L(Pc, Pb, Qb, 4),  AI_F(Pc, Pb, Qc, Qb, 4);
                                  1.213,  1.252,  1.193,  1.203
```

**Рис. 32. Вычисление агрегатных индексов цен в форме *Пааше*, *Ласпейреса* и *Фишера***

На основе указанных функций с использованием данных табл. **20** мы получаем значения агрегатных индексов цен, а именно: **AI_P1(Pc, Pb, Qc ,4) =** 1.213 (*тип Пааше* **1**), **AI_P2(Pc, Pb, Qc, Qb, 4)** = 1.252 (*тип Пааше* **2**) и **AI_L(Pc, Pb, Qb, 4)** = 1.193 (*тип Ласпейреса*). Естественно, в случае других реальных данных эти различия в значениях агрегатных индексов могут быть намного более существенными.

Таким образом, *индекс Ласпейреса* говорит нам о том, что в среднем вышеупомянутый список (табл. **20**) стоил бы в 2000 г. 1.193*100 = 119.3% относительно стоимости 1995 г. Необходимо обратить внимание, что форма Ласпейреса взвешенного агрегатного индекса – основная теоретическая модель для нескольких ценовых индексов, издаваемых Департаментом труда США, хотя эквивалентные взвешенные средние обычно используются для целей реальных вычислений. Наконец, результат, полученный на основе *индекса Пааш* первой формы говорит нам, что в среднем *текущий* список (2000) стоил бы 1.252*100 = 125.2% относительно *основного* периода (1995).

И. Фишер на основе вышеуказанных взвешенных агрегатных индексов Пааше и Ласпейреса ввел так называемый ***идеальный индекс*** (**AI_F**), который является попросту геометрическим средним двух взвешенных агрегатных индексов, а именно:

$$AI\_F = \sqrt{Paasche's\ index * Laspeyres's\ index}$$

Для иллюстративных данных (табл. **20**), которые мы использовали выше, легко получаем индекс *Фишера* **AI_F(Pc, Pb, Qc, Qb, 4)** = 1.203 (рис. **32**). Как увидим позже, индекс И. Фишера "*идеален*" в том смысле, что он удовлетворяет некоторым логическим критериями хорошего индексного показателя. Никакой другой индекс не удовлетворяетт всем этим критериям. Однако, индекс Фишера, как правило, затруднителен для использования из-за практических соображений. Между тем, индекс среднего геометрического Фишера достаточно широко используется в ряде стран. Тогда как в отечественной статистике этот индекс используется достаточно редко, прежде всего, в международных сравнениях.

Однако, выбор типа *агрегатного индекса* должен, прежде всего, основываться на результатах *анализа* явления, характер изменения которого он должен отражать. Так, для оценки научной деятельности (*значимости*) **ТТГ** нами введен ряд индексов. Индексы, рассмвтриваемые ниже, могут использоваться для оценок творческой деятельности научных коллективов, включая другие творческие группы и отдельных лиц. Так, ***индекс востребованных работ*** (**AI_CW**) вычисляется по следующей формуле типа (75.2):

$$\mathbf{AI\_CW} = \frac{\sum_{k=1}^{n} \mathbf{Qc_k\ Wc_k}}{\sum_{k=1}^{n} \mathbf{Qb_k\ Wb_k}} \qquad (77)$$

где **W** – объемы цитированных работ и **Q** – количества ссылок на них, исключая авторские ссылки. Изменение величины **AI_CW** зависит как от количества цитированных работ, так и от их объемов, а во времени она характеризует изменение значимости (*востребованности*) научной продукции относительно *предыдущего* или некоторого *базисного* периода. Например,

*востребованность* работ **ТТГ** в 1999 г. возросла более, чем в 16 раз относительно 1971 г. (*взятого в качестве базисного, как начала цитирования работ* **ТТГ;** *см. табл.* **9**).

Индивидуальный индекс ***фундаментальности работы*** (**II_WS**) вычисляется по следующей простой формуле:

$$\mathbf{II\_WS = (Tc*Qc + 1) / Tb*Qb} \tag{78}$$

где **Tc –** длительность существования публикации до текущего периода (*в годах*), **Qc –** число ссылок на нее, исключая авторские, в текущем году; **Tb = Qb = 1 –** исходные данные *базисного* периода – года первой публикации работы. Если публикация никогда или в текущем году не цитировалась, то для нее полагается **II_WS = 1,**, что отражает только сам факт публикации, в противном случае его величина растет с ростом периода активной жизни и цитируемости публикации, что, вообще говоря, можно определять как уровень ее фундаментальности. Так, значения индекса **II_WS** в 1999 г. для наших монографий [1, 4, 12] соответственно составили: (27*58 + 1) = 1567, (19*125 + 1) = 2376 и (9*62 + 1) = 559, что на указанный период говорит нам о существенно большей фундаментальности монографии [4]. Для значение показателя **II_WS** немалую роль сыграла *англоязычность* публикации. Тогда ***индекс общей фундаментальности публикаций*** (**II_WS**) вычисляется по следующей формуле второго типа Пааше (75.2):

$$\mathbf{AI\_CS} = \frac{\sum_{k=1}^{n} \mathbf{II\_WSc_k\, Wc_k}}{\sum_{k=1}^{n} \mathbf{II\_WSb_k\, Wb_k}} \tag{79}$$

где **II_WSk –** индивидуальный индекс фундаментальности **k** -публикации и **Wk –** ее объем (**k = 1 .. n**). Вообще говоря, для характеристики динамики различных явлений могут создаваться самые разнообразные индексы и их системы, основным требованием к которым является адекватность отражения специфики изучаемого процесса. Выбор математической формулы индекса, его весов и т.п. определяется характером описываемых им явлений. Разнообразие исследуемых явлений, естественно, приводит к разнообразию измеряющих их методов и приемов. Общие принципы построения экономических индексов, а также индексов иного назначения можно найти в интересных книгах [64-67, 83, 92, 96, 114-116, 122, 124, 125, 209].

## 9.3. Средние, цепные и базисные индексы

Кроме *агрегатных* и *индивидуальных* в статистике и экономике применяются также *средние* индексы, вычисляемые как *средние* величины из индивидуальных индексов. Вычисляются индексы ***средний взвешенный*** и ***невзвешенный*****,** среди которых наиболее употребительны средне арифметические и средне гармонические. В ряде стран широко используется средне геометрический индекс И. Фишера – средне геометрическое из произведения индексов типа Пааше и Ласпейреса. В отечественной же статистике он используется весьма редко, в первую очередь, при разных *международных* сопоставлениях. ***Средний арифметический индекс*** (**AAI**) вычисляется как средне взвешенная арифметическая из индивидуальных индексов (**ИИ**). В отечественной статистике **AAI** вычисляется как агрегатный индекс путем преобразования последнего, состоящего в замене индексируемой величины *текущего* периода *произведениями* значений **ИИ** и значений индексируемой величины *базисного* периода. Так, если **AAI** имеет общую формулу типа Пааше (75.1), то **ИИ** принимает вид **II(X) = Xc / Xb,** т.е. **Xc = II(X)*Xb,** откуда формула **AAI** принимает следующий общий вид:

$$\mathbf{AAI(X)} = \frac{\sum_{k=1}^{n} \mathbf{II(X)_k\, Xb_k\, Wc_k}}{\sum_{k=1}^{n} \mathbf{Xb_k\, Wc_k}} \tag{80}$$

где компоненты **Xb** и **Wc** имеют тот же самый смысл, что и в случае формул (75). Из формулы (80) следует, что взвешивание **ИИ** производится произведениями соответствующих значений (*в зависимости от конкретного содержания индекса*) *индексируемой* **X**-величины *базисного* периода и значений показателя, служащих *весами* в агрегатном индексе. Показатель **AAI** применяется в тех случаях, когда прямое использование индексируемой величины текущего периода в агрегатном индексе (**АИ**) встречает какие-либо затруднения.

***Средне гармонический индекс*** (**AHI**) вычисляется как средне взвешенная гармоническая из **ИИ**. В отечественной статистике **AHI** вычисляется как **АИ** путем преобразования последнего, состоящего в замене индексируемой **X**-величины базисного периода отношениями значений индексируемой величины текущего периода к значениям **ИИ**. Так, если **АИ** имеет общую формулу первого типа Пааше (75.1), то **ИИ** принимает вид **II(X) = Xc / Xb,** т. е. **Xb = Xc / II(X),** откуда формула **AHI** принимает следующий общий вид:

$$\mathbf{AHI(X)} = \frac{\sum_{k=1}^{n} \mathbf{Xc_k\, Wc_k}}{\sum_{k=1}^{n} \frac{\mathbf{Xc_k\, Wc_k}}{\mathbf{II(X)_k}}} \qquad (81)$$

где компоненты **Xc** и **Wc** имеют тот же самый смысл, что и в случае формул (75). Из формулы (81) следует, что взвешивание производится произведениями соответствующих значений (*в зависимости от конкретного содержания индекса*) индексируемой **X**-величины *текущего* периода и значений показателя, служащих в **АИ** весами. Показатель **AHI** используется в тех случаях, когда прямое использование базисной индексируемой величины в **АИ** встречает какие-либо затруднения либо его применение более целесообразно с вычислительной точки зрения.

***Средне невзвешенный индекс*** (**AUI**) вычисляется по формуле простой средней, а именно:

$$\mathbf{AUI(X)} = \frac{\sum_{k=1}^{n} \mathbf{II(X)_k}}{\mathbf{n}}$$

и в практике отечественной статистики не применяется. Как нетрудно убедиться из простых алгебраических преобразований формул (75.1, 80, 81) и соотношения **II(X)=Xc / Xb,** для **ИИ** имеет место общее тождество, а именно: **AI(X) ≡ AAI(X) ≡ AHI(X),** что позволяет выбирать метод вычисления **АИ** в зависимости от конкретных условий. В качестве довольно полезного упражнения читателю рекомендуется получить формулы средних индексов относительно агрегатных индексов всех трех типов (75).

В качестве примера мы вычислим вышеупомянутые средние индексы относительно данных табл. **20**. Прежде всего, мы вычисляем значение индивидуального индекса (**II**) для каждого вида товаров **G, S, Art** и **Kr,** а именно: **II(G)** = 53/48 = 1.1, **II(S)** = 33/28 = 1.2, **II(Art)** = 10/6 = 1.7 и **II(Kr)** = 4/2 = 2.

| | | |
|---|---|---|
| **AAI(X)=** | (1.1*48*942+1.2*28*995+1.7*6*998+2*2*2005)/(48*942+28*995+6*998+2*2005) | **=1.22** |
| **AHI(X)=** | (53*942+33*995+10*998+4*1005)/(53*942/1.1+33*995/1.2+10*998/1.7+4*1005/2) | **=1.20** |
| **AUI(X)=** | (1.1+1.2+1.7+2)/4 | **=1.50** |

Как следует из расчетов, значения первых двух типов *средних индексов* (*с точностью до второго десятичного знака*) совпадают с ранее расчитанным значением совокупного индекса в форме Пааше *первого* типа. Кроме того, имеет место несоответствие между значениями **AUI,** с одной стороны, **AAI** и **AHI,** с другой стороны. Однако, при ином *распределении* значений начальных данных, ситуация может существенно измениться относительно расхождения значений **AUI,** с одной стороны, и **AI, AAI** и **AHI,** с другой стороны.

Для того же самого примера (табл. **20**) вычислим ***средний геометрический индекс*** (**AGI**) *цен*, взвешенных согласно товарообороту *текущего* периода (*для упрощения вычислений формула представлена в логарифмическом виде*):

$$\ln(\mathbf{AGI}) = \frac{\sum_{k=1}^{n} (\ln(Pc_k) - \ln(Pb_k))\, Pc_k Qc_k}{\sum_{k=1}^{n} Pc_k Qc_k}$$

[{ln(53) - ln(48)}*53*942 + {ln(33) - ln(28)}*33*995 + {ln(10) - ln(6)}*10*998 + {ln(4) - ln(2)}*4*2005]/ (53*942 + 33*995 + 10*998 + 4*2005) = 0.208; **AGI = 1.23**

Вышеприведенное вычисление посредством формулы для **AGI** дает значение достаточно близкое к ранее полученным значениям показателей **AAI** и **AHI** (1.22 и 1.20 соответственно), однако процесс вычислений становится более трудоемким.

В целях более адекватного отражения исследуемого явления используются *системы индексов*, представляющие собой последовательные ряды индексов с *переменной* или *постоянной базами* сравнения, с переменными или постоянными весами, а также целый ряд взаимосвязанных по экономическому содержанию индексов. Например, система индексов *цен*, *физического объема* товарооборота и *товарооборота* в ценах соответствующих периодов. При построении системы индексов необходимо строго соблюдать взаимосвязь используемых ею весов. Мы рассмотрим системы *базисных* и *цепных* индексов.

***Базисные индексы*** (**BI**) представляют собой систему индексов одного и того же явления с постоянным базисом сравнения, т.е. в знаменателях всех выражений индексов используется индексируемая величина *базисного периода*. **BI** могут быть как *индивидуальными*, так и *общими*. В свою очередь, *общие* **BI** могут быть как с *постоянными*, так и с *переменными* весами. Так, если **X** – индексируемая величина и **W** – ее веса, то *общие* формулы для **BI** *индивидуальных* и *общих* с *постоянными* и *переменными* весами представлены в табл. **21** (строки **1, 2** и **3** соответственно). Выбор весов индексов определяется на основе логического анализа исследуемого явления.

**Таблица 21. Формулы для базисных и цепных индексов**

| No | *Формулы* | *Тип индекса* |
|---|---|---|
| **1** | $\mathbf{BI_1} = \frac{X_1}{Xb}, \ldots, \mathbf{BI_k} = \frac{X_k}{Xb}, \ldots, \mathbf{BI_n} = \frac{X_n}{Xb}$ | *individual* **BI** |
| **2** | $\mathbf{BI_1} = \frac{\sum_{k=1}^{n} X1_k W_k}{\sum_{k=1}^{n} Xb_k W_k}, \ldots, \mathbf{BI_n} = \frac{\sum_{k=1}^{n} Xn_k W_k}{\sum_{k=1}^{n} Xb_k W_k}$ | **BI** *with constant weights* |
| **3** | $\mathbf{BI_1} = \frac{\sum_{k=1}^{n} X1_k W1_k}{\sum_{k=1}^{n} Xb_k W1_k}, \ldots, \mathbf{BI_n} = \frac{\sum_{k=1}^{n} Xn_k Wn_k}{\sum_{k=1}^{n} Xb_k Wn_k}$ | **BI** *с переменными весами* |
| **4** | $\mathbf{CI_1} = \frac{X_1}{Xb}, \ldots, \mathbf{CI_k} = \frac{X_k}{Xk-1}, \ldots, \mathbf{CI_n} = \frac{X_n}{Xn-1}$ | *индивидуальные* **CI** |

| | | |
|---|---|---|
| **5** | $\mathbf{CI_1 = \frac{\sum_{k=1}^{n} X1_k W_k}{\sum_{k=1}^{n} Xb_k W_k}, \ldots, CI_n = \frac{\sum_{k=1}^{n} Xn_k W_k}{\sum_{k=1}^{n} Xb_{n-1_k} W_k}}$ | **CI** *с постоянными весами* |
| **6** | $\mathbf{CI_1 = \frac{\sum_{k=1}^{n} X1_k W1_k}{\sum_{k=1}^{n} Xb_k W1_k}, \ldots, CI_n = \frac{\sum_{k=1}^{n} Xn_k Wn_k}{\sum_{k=1}^{n} Xb_{n-1_k} Wn_k}}$ | **CI** *с переменными весами* |
| | X – *индексированная величина*, W – *ее веса и* Xb – *значения базисного уровня* | |

***Цепные индексы*** (**CI**) представляют собой также систему индексов одного и того же явления с *переменным базисом* сравнения. **CI** подобно **BI** могут быть как *индивидуальными*, так и *общими*. В свою очередь, *общие* **CI** могут быть как с *постоянными*, так и с *переменными* весами. Общие формулы для **CI** *индивидуальных* и *общих* с *постоянными* и *переменными* весами представлены в табл. **21** (строки **4, 5** и **6** соответственно). Между **BI** и **CI** существует очевидная взаимосвязь: (**1**) *произведение* **k** последовательных **CI** системы равно соответствующему **k**-му **BI**; (**2**) *отношение* текущего **BI** к предшествующему дает соответствующий текущий **CI**. Данная взаимосвязь всегда имеет место для *индивидуальных* индексов, а для *общих* только при *постоянстве* весов или их соизмерителей. Читателю рекомендуется проверить справедливость сказанного на общих формулах индексов **CI** и **BI** (табл. **21**).

***Общие индексы*** по одному виду продукции, относящейся к различным объектам, могут вычисляться двумя способами, а именно: как индексы (**1**) *фиксированного* или (**2**) *переменного* состава. Индексы *фиксированного* состава являются **АИ,** тогда как индексы *переменного* состава вычисляются сопоставлением средних по всем объектам уровней *индексируемых* показателей, например, сопоставлением *средней* по всем предприятиям *себестоимости единицы* продукции в текущем периоде со средней себестоимостью в базисном периоде. Отношением индекса переменного состава к индексу фиксированного состава можно определять так называемый ***индекс структурных сдвигов*** (**ISP**), в целом характеризующий *влияние* изменения структуры совокупности. Необходимым условием построения индексов переменного состава служит возможность вычисления среднего уровня. Поэтому, такого типа индексы можно строить по таким качественным показателям, как зарплата, себестоимость, производительность труда, цитируемость публикаций и др. Проиллюстрируем как их построение, так и особенности на примере показателя цитируемости научных публикаций (**ISP**). Данный индекс построен с ориентацией на наши собственные нужды анализа динамики публикации научных работ **ТТГ,** однако показатель может использоваться и более широко.

***Средняя цитируемость публикаций*** за базисный год выражается как средне взвешенная по типам публикаций (*теоретические работы по* **МТОС – k = 1,** *работы по прикладным аспектам* **МТОС – k = 2,** *другие работы* – **k = 3**) за *базисный* период сравнения:

$$\mathbf{Qb^{Sr} = \frac{\sum_{k=1}^{3} Qb_k}{\sum_{k=2}^{3} Nb_k}} \qquad (82)$$

тогда как за текущий период (*год*) выражается следующей формулой:

$$\mathbf{Qc}^{\mathbf{Sr}} = \frac{\sum_{k=1}^{3} \mathbf{Qc_k}}{\sum_{k=2}^{3} \mathbf{Nc_k}} \tag{83}$$

где для обоих формул (82, 83): $\mathbf{Qc_k}$ ($\mathbf{Qb_k}$) и $\mathbf{Nc_k}$ ($\mathbf{Nb_k}$) – соответственно *цитируемость* и *число* публикаций **k**-типа в *текущем* (*базисном*) году соответственно.

Для определения *степени изменения цитируемости* находим отношение $\mathbf{AIvar = Qc^{Sr} / Qb^{Sr}}$, показывающее как степень изменения цитируемости, так и степень изменения структуры публикаций. Показатель **AIvar –** *индекс цитируемости переменного состава*. Но для нахождения *изменения* самой цитируемости нужно построить *индекс постоянного состава* (**I*const***), в котором взвешивание будет вестись по типам публикаций текущего периода, то есть по следующей общей формуле:

$$\mathbf{I}\boldsymbol{const} = \frac{\sum_{k=1}^{3} \mathbf{Qc_k\, Nc_k}}{\sum_{k=1}^{3} \mathbf{Qb_k\, Nc_k}} \tag{84}$$

Тогда для определения изменения структуры публикаций (**ISP**) по приведенному выше соотношению нужно вычислить отношение $\mathbf{I}\boldsymbol{cit} = \mathbf{Aivar/I}\boldsymbol{const}$, в общем виде [*после проведения простых преобразований с учетом формул* (82 - 84)] представляемому следующим образом:

$$\mathbf{I}\boldsymbol{cit} = \frac{\sum_{k=1}^{3} \mathbf{Qb_k\, Nc_k} \sum_{k=1}^{3} \mathbf{Nb_k}}{\sum_{k=1}^{3} \mathbf{Qb_k\, Nb_k} \sum_{k=1}^{3} \mathbf{Nc_k}} \tag{85}$$

Следовательно получаем, **ISP** равен *агрегатному индексу* типов публикаций, *взвешенному* по их *цитируемости* в *базисном* году, деленному на *индекс изменения количества публикаций*. Каждый из индексов, составляющих **ISP,** решает определенную задачу и применим на разных *уровнях анализа* и *управления*: **AIvar** характеризует *общее* изменение размеров *качественного* показателя, в нашем примере средней цитируемости публикаций **ТТГ; I*const*** показывает, как изменился уровень *качественного* показателя независимо от изменения структуры совокупности (*состава публикаций*). Подобным же образом можно строить *индексы изменения пропорций* (**ISP)** любых качественных показателей, уровень которых может выражаться *средними* величинами.

## 9.4. Важнейшие экономические индексы и их взаимосвязь

Рассмотренные общие теоретические положения индексного метода позволяют прояснить особенности вычисления *индексов* отдельных *экономических* показателей. Кратко рассмотрим важнейшие из них, а также имеющиеся взаимосвязи между ними. На использовании таких взаимосвязей индексов основано разложение *сложных* экономических индексов на множество составляющих их элементов.

***Индексы цен*** (**ИЦ**) – обобщающие показатели динамики и соотношения уровня цен; могут быть *индивидуальными*, определяемыми для отдельных видов товаров или услуг, и *сводными*, характеризующими соотношение уровня цен по совокупности различных товаров и услуг. *Сводные* **ИЦ,** в свою очередь, делятся на *общие*, охватывающие всю изучаемую совокупность, и *групповые*, определяемые для *отдельных групп товаров*. *Сводные* **ИЦ** вычисляются по формулам *агрегатных индексов* и *средне гармонических индексов;* обе формулы для **ИЦ** дают идентичные

результаты и их выбор определяется характером исходных данных. Формула *агрегатных индексов* применяется, если известны *цены* и *количества отдельных* видов товаров в *натуральном* выражении. Тогда как формула *средне гармонических* индексов используется, если известны индивидуальные индексы цен по отдельным видам или группам товаров и стоимостные выражения товаров.

Согласно современной классификации имеется пять основных типов индексов цен. Они делятся на два класса, а именно: *простые* индексы и *взвешенные* индексы. ***Простые*** индексы важны в качестве помощи при разработке задачи создания индекса цен и представления символов, используемых в формулах, но вне этого, они не имеют каких-либо существенных применений. В свою очередь, пять типов индексов цен могут быть классифицированы как:

| | | |
|---|---|---|
| **1** | ***Простые агрегатные индексы*** | ***Простые индексы*** |
| **2** | ***Простая средняя ценовых отношений*** | |
| **3** | ***Взвешенные агрегатные индексы:*** **(a)** ***форма Пааше,*** **(b)** ***форма Ласпейреса*** | ***Взвешенные индексы*** |
| **4** | ***Идеальный индекс Фишера*** | |
| **5** | ***Всзвешенные средние ценовых отношений*** | |

Эти типы индексов рассматривались в предыдущем разделе независимо от отражаемой ими сущности. С точки же зрения практических приложений наиболее важными представляются *взвешенные индексы*. Форма *идеального* индекса Фишера используется в сфере международной торговли, однако в других приложениях она широко не используется, в значительной мере потому, что некоторые из требуемых расчетных данных не могут быть получены в каждый данный период. Таким образом, наиболее применяемыми на практике являются *взвешенные индексы* типов (**3**) и (**5**).

**ИЦ** вычисляются для *отдельных* видов цен, отражая динамику их уровня в различных сферах товарно-денежного обращения. Основными видами здесь являются индексы оптовых цен на промышленную продукцию, государственных розничных цен, цен рыночной торговли, закупочных цен, цен на строительную продукцию, цен и тарифов на услуги, цен внешней торговли. Для характеристики соотношения уровня цен в различных районах страны или при международных сопоставлениях вычисляются территориальные индексы цен.

***Сводные индексы цен*** (**I_P**) вычисляются на основе формул *агрегатных индексов* или *средних гармонических индексов*; они выражаются посредством следующих двух формул:

$$\textbf{(a)}\quad \mathbf{I\_P} = \frac{\sum_{k=1}^{n} \mathbf{Pc_k}\,\mathbf{Qc_k}}{\sum_{k=1}^{n} \mathbf{Pb_k}\,\mathbf{Qc_k}} \qquad \textbf{(b)}\quad \mathbf{I\_P} = \frac{\sum_{k=1}^{n} \mathbf{Pc_k}\,\mathbf{Qc_k}}{\sum_{k=1}^{n} \frac{\mathbf{Pc_k}\,\mathbf{Qc_k}}{\mathbf{Ip_k}}} \tag{86}$$

где **Pk –** цена, **Qk –** количество товаров, и $\mathbf{Ip_k}$ **–** индивидуальный индекс цены **k**-х товаров.

В дальнейшем как и прежде знаки "**b**" и "**c**" обозначают соответственно *базисные* и *текущие* периоды. Обе формулы для значений индексов цен дают идентичные результаты и их выбор определяется, прежде всего, характером начальных данных. При этом, формула *агрегатного индекса* (**86.a**) используется, если цены и количества отдельных видов товаров известны в реальном исчислении. Тогда как формула (**86.b**) *среднего гармонического индекса* используется, если известны ***индивидуальные индексы цен*** (**Ip**) по отдельным видам или группам товаров наряду с их стоимостным выражением.

***Индексы розничных цен*** (**CPI**) играют чрезвычайно важную роль для мониторинга уровня жизни населения. Исходной формой **CPI** является агрегатный индекс типа (**86.a**), однако ее

использование встречает серьезные затруднения, ибо в ряде важных отраслей торговли первичный учет строится на стоимостном, а не на количественном принципе. В этом случае используется эквивалентная ей форма средне гармонического индекса (**86.b**), алгебраически тождественная формуле Пааше и имеющая такое же экономическое содержание. С **CPI** тесно связан *показатель реальной зарплаты*, для мониторинга которого они часто используются. В США **CPI** вычисляются и издаются *ежемесячно*. **CPI** вместе с другими данными используются для построения других важных индексов типа *розничного ценового индекса* и *дефлятора валового национального продукта*.

***Индекс оптовых цен*** (**WPI**) определяется ценами, по которым производители продают на *первичных* рынках, а не ценами, по которым продают *оптовые* торговцы. Все виды предметов потребления от сырья до изготовленных изделий оцениваются при построении индекса. Большинство ценовых данных по изготовленной продукции поступает непосредственно от производителей. Основная формула для вычисления этого индекса – модифицированная взвешенная средняя ценовых отношений с используемыми весовыми значениями. **WPI** рассчитывается и издается ежемесячно американским Бюро статистики труда. В дополнение к ежемесячному **WPI,** издаются *еженедельный* и *ежедневный* индексы. Важное использование **WPI** включает в себя регулирование денежных сумм для изменений цен в долгосрочных контрактах различных типов. **CPI** и **WPI** – одни из ведущих ценовых индексов, которые оба издаются американским Бюро статистики труда.

***Индексы себестоимости продукции*** (**I_PC**) получили весьма широкое применение в *практике* отечественного учета и планирования. Принципы их построения такие же, что и для **ИЦ.** Экономический смысл **I_PC** состоит в том, что они выражают результаты хозяйственной деятельности предприятий различного типа путем сопоставления фактических затрат на производство продукции с (*нормальными*) нормативными затратами. Так как учет внутри предприятия всегда дает возможность *прямо* или *косвенно* определять *себестоимость* единицы отдельного вида продукции, то для вычисления **I_PC** используется формула типа Пааше, а именно формула (**87.a**):

$$\textbf{(a)}\ \ \mathbf{I\_PC} = \frac{\sum_{k=1}^{n} \mathbf{Zc_k Qc_k}}{\sum_{k=1}^{n} \mathbf{Zb_k Qc_k}} \qquad \textbf{(b)}\ \ \mathbf{I\_PC} = \frac{\sum_{k=1}^{n} \mathbf{Zc_k Qc_k}}{\sum_{k=1}^{n} \frac{\mathbf{Zc_k Qc_k}}{\mathbf{Iz_k}}} \tag{87}$$

в которой $\mathbf{Z_k}$ – себестоимость единицы **k**-продукции и $\mathbf{Q_k}$ – ее количество; знаки "**b**" и "**c**" обозначают соответственно базисный и текущий периоды. Аналогично случаю **ИЦ** для вычисления **I_PC** используется форма средне гармонической (**87.b**), в которой $\mathbf{Iz_k}$ – **ИИ** себестоимости **k**-продукции при (**k = 1 .. n**).

По сравнению с **CPI I_PC** вычисляется всегда по *строго определенной номенклатуре* продукции. В отраслях производства с быстрым и существенным изменением номенклатуры продукции возникает несопоставимость между **I_PC** и другими индексами, относящимися ко всей *совокупности* продукции. Поэтому, *как правило*, индексы **I_PC** вычисляются для двух смежных календарных лет. При необходимости вычислять сопоставления по более продолжительным интервалам времени (*например*, *пятилеткам*) использование **I_PC** в ряде случаев приводит к неверным выводам. Поэтому для этих целей используются другие показатели, например, *индекс удельного веса себестоимости* в валовой или товарной продукции, который представляет собой разновидность индексов с переменным составом. Однако и здесь возникают сложности в системе оценки продукции.

***Индексы*** физического ***объема продукции*** (**VIP**) наиболее яркое выражение в отечественной статистике находят в промышленности, выступая в форме *индекса промышленной продукции* (**IIP**). В этом случае в основу берется отношение сопоставимых показателей объема всей продукции в *текущем* и *базисном* годах. **IIP** вычисляется по формуле типа Ласпейреса (**88.a**):

$$\textbf{(a)}\ \ I\_IIP = \frac{\sum_{k=1}^{n} Pb_k\, Qc_k}{\sum_{k=1}^{n} Pb_k\, Qb_k} \qquad \textbf{(b)}\ \ I\_IIP = \frac{\sum_{k=1}^{n} Iq_k\, Pb_k\, Qb_k}{\sum_{k=1}^{n} Pb_k\, Qb_k} \tag{88}$$

где некоторая *базисная цена* (**Pb**) не является единственно возможным *измерителем продукции*. Используя **ИИ** объема продукции **Iq = Qc / Qb,** данная формула получает эквивалентный вид, как средне взвешенная арифметическая величина (**88.b**). Теоретически более удобным измерителем являлись бы базисные затраты на единицу продукции, однако здесь возникают серьезные практические сложности при вычислениях этих затрат.

*Особенностью* определения **IIP** в СССР было то, что он являлся *результатом непосредственного суммирования* отчетных *данных* предприятий. Это объяснялось тем, что *каждое* промышленное предприятие оценивало свою продукцию как в *текущих*, так и в *базисных* ценах, ибо *плановые* задания совокупного объема промышленной продукции определялись именно в базисных ценах. Важным методологическим вопросом здесь является выбор самой базисной цены. Для измерения динамики продукции в качестве последних базисных цен были выбраны цены 1975 г. Принципы построения индекса физического объема сельхозпродукции аналогичны сказанному относительно **IIP**. Однако его вычисление производилось непосредственно ЦСУ, ибо сельхозпредприятия не вели двойной оценки (*в текущих и базисных ценах*) продукции. Однако сравнительно с промышленностью существенно более ограниченная номенклатура сельхозпродукции делает возможным производить *централизованную* оценку всего ее объема в сопоставимых ценах.

В США **IIP** является широко распространяемым деловым индикатором, показывающим относительные изменения в физических объемах продукции. Он имеет базовое значение для отраслей промышленности, охваченных им – производство, горная промышленность, электрические и газовые предприятия. Вероятностные выборочные методы не используются. Данный индекс строится на данных, собранных изначально для других целей. Эти данные предоставляются как *частными* деловыми источниками, так и *правительственными* органами. Вместе с другими индикаторами, **IIP** используется, чтобы установить состояние совокупной экономики. Анализ индивидуальных индексов производства промышленности указывает те сектора, которые вносят наибольший вклад в общий подъем или спад производства, если это случается. При этом, *индивидуальный бизнес* может измерить *собственную* работу производства, анализируя соответствующие данные индекса производства.

***Индекс физического объема товарооборота*** (**VIT**) определяется как отношение стоимости *товарного объема* к *индексу цен* (**PI**), т.е. для промышленности, например, **VIT=IIP/PI**. Следует, однако, отметить, что особенности современной организации статистики товарооборота в *государственной* и *иных* формах торговли не дают возможности проводить непосредственные измерения в сопоставимых ценах.

***Индексы производительности труда*** (**ILP**) играют весьма важную роль в анализе динамики ***производительности труда*** (**LP**). Однако здесь *до сих пор* имеется много спорных и открытых вопросов. Трудности данного анализа, в частности, связаны с тем, что для экономической категории **LP** не всегда можно подобрать адекватную статистическую интерпретацию. В общем виде уровень **LP** представляет собой *отношение объема продукции* к *затраченному* на ее

изготовление *труду*. Разные содержания, вкладываемые в числитель и/или знаменатель данной формулы, породили много вариантов показателей **LP**. Теоретически доказано, что наилучшей для **ILP** является формула (**89.a**), а именно:

$$\textbf{(a)}\ \ I\_ILP = \frac{\sum_{k=1}^{n} Tb_k\, Qc_k}{\sum_{k=1}^{n} Tc_k\, Qc_k} \qquad \textbf{(b)}\ \ I\_ILP = \frac{\sum_{k=1}^{n} Iw_k\, Tc_k\, Qc_k}{\sum_{k=1}^{n} Tc_k\, Qc_k} \tag{89}$$

в которой **Tk** – трудозатраты на единицу **k**-продукции, или эквивалентная ей формула (**89.b**) средне взвешенной арифметической, если использовать подстановку в первую формулу **ИИ** *трудозатрат* следующего вида **Iw = Tb / Tc.** Однако, из-за трудности учета об универсальном применении этих формул **ILP** говорить не приходится, что привело к значительному числу различного типа форм **ILP**.

В частности, нами определен ***индекс среднемесячной творческой активности*** (**IAMCA**), используемый для оценки деятельности **ТТГ,** характеризуемой ее публикациями различного типа. Суть данного показателя состоит в следующем. Каждая **k**-публикация (*как законченная работа или ее этап*) характеризуется *объемом* (**Vk**) в печатных листах (**1** *п.л.* = 40.000 *символов*) и *трудозатратами* (**Lk**) различного рода (*интеллектуальными*, *техническими и др.*), *обобщаемыми* в виде длительности разработки в месяцах. При довольно хорошо налаженной в **ТТГ** системе координации работ, *текущего* и *перспективного* планирования работ, а также использовании компьютерного набора рукописей нетрудно оценивать указанные показатели с достаточно удовлетворительной степенью точности (*первый – весьма точно по данным компьютерного набора*, *второй – с точностью до недели*, *т.е.* 1/4 *месяца*). В текущем году в качестве учетных единиц выбираются те, работа над которыми была полностью проведена в рамках данного года (*начата и доведена до уровня "подготовлена к публикации"* (*Submitted for publication*) или передана в печать, а также *переходящие* работы от предыдущих лет и завершенные в *текущем*. Переходящие работы учитываются в смежных годах пропорционально трудозатратам на них. Данный подход требует постепенной корректировки данных с учетом завершения переходящих работ. Для статистического анализа деятельности **ТТГ** нами использовались два основных показателя активности: *индивидуальный индекс активности* (**IIA**) и **IAMCA,** вычисляемые по следующим простым формулам, а именно:

$$IIA_k = \frac{V_k}{L_k} \qquad IAMCA = \frac{1}{12}\left(\sum_{k=1}^{n} V_k\right)$$

*Первый* показатель характеризует творческую активность по отношению к *конкретной* работе (*здесь на сегодня лидерами являются наши книги* [1, 2, 36, 29, 140, 141] *и* [143] *со значениями* **ИИА** соответственно 3.1, 3.4, 3.8, 4.0, 4.8, 6.3 и 7.1), *второй* – среднемесячную активность в текущем году (*наиболее активными годами деятельности* **ТТГ** *здесь являются* 1997, 1998 *и* 1999, *для которых значения* **IAMCA** соответственно равняются 105.3, 140.5 и 190.8). На основе второго показателя легко строится соответствующий *агрегатный индекс* в форме Ласпейреса (**75.2**). Оставляет это читателю в качестве полезного упражнения. Еще раз необходимо обратить внимание, что на сегодня представленные выше статистические данные об активности **ТТГ** существенно иные. Однако, это не совершенно не отражается на их качестве как иллюстративного материала.

Как уже отмечалось выше, ***индекс*** представляет собой относительную величину, получаемую в результате сопоставления уровней социально-экономических явлений с течением времени, в пространстве либо с плановыми показателями. В качестве меры соизмерения разнородных продуктов можно использовать цену, себестоимость или трудоемкость единицы продукции.

В развитии индексной теории в СССР сложились два основных направления: *обобщающее* или *синтетическое*, и *аналитическое*.

Различие между этими направлениями обусловлено двумя возможностями интерпретации *индексов* в их прикладном аспекте, а именно. *Обобщающее* или так называемое ***синтетическое*** направление трактует *индекс* как показатель среднего изменения уровня изучаемого явления. Тогда как в ***аналитической*** теории *индексы* воспринимаются как показатели *изменения уровня* результативной величины под влиянием изменения индексируемой величины. Развитие второго направления было обусловлено применением индексного метода в экономическом анализе. При этом, способы построения индексов зависят от содержания изучаемых явлений, методологии расчета исходных статистических показателей и самих целей исследования.

Сложность социально-экономических явлений, наличие многосторонних связей между ними приводит к тому, что отдельно взятый показатель не в состоянии сколько-нибудь полно отражать все богатство *содержания* этих явлений. Следовательно, отпадает идея *универсальных* показателей, ибо каждый отдельный показатель отражает только одну из сторон динамики явлений или одну из сторон *взаимосвязи* между ними. Только система индексов и показателей позволяет всесторонне отражать динамику явлений в их совокупности. В масштабе отрасли или всего народного хозяйства взаимосвязь индексов и показателей имеет особое значение, что требует разработки тщательно продуманной системы взаимосвязанных индексов. Речь здесь идет не о том, чтобы каждый индекс удовлетворял одним и тем же формальным требованиям, например, тестам И. Фишера, а о том, чтобы взаимосвязь *индексов* и *показателей* выражала реально существующие экономические взаимоотношения между явлениями и целыми отраслями. Указанная взаимосвязь позволяет измерять факторы, определяющие динамику того или иного явления. Так, объем произведенной продукции определяется как функция двух величин: производительности труда и затраченного рабочего времени. Таким образом, методы построения индексов продукции и индексов производительности труда должны взаимоувязывать оба указанных экономических индекса.

Наконец, разработка различных видов тестов на допустимость тех или индексов касается важной проблемы взаимосвязи индексов. Все это требует создания некоторых логических методов проверки индексов и их систем на предмет адекватности отражения ими динамики изучаемых явлений. Ряд этих вопросов будет рассматриваться в следующем разделе главы.

## 9.5. Логические критерии хороших индексов

В предыдущем разделе мы вкратце познакомились с *индексами*. Мы увидели, что в практике индексного анализа приходится вводить некоторые допущения, не соответствующие логике теории. Историческое развитие науки об индексах складывалось в борьбе разных позиций по поводу оправданности подобных допущений с точки зрения практического применения индексов. Это вполне естественно, т.к. индексы зародились и развивались, прежде всего, как инструмент прикладного анализа. Между тем, далеко не сразу практическая значимость индексов была признана широкими статистическими и экономическими кругами. Дело в том, что *индекс* можно вычислять по разным формулам, что приводит к разным результатам.

Так как для построения индексов имеется много подходов, дающих разные вычислительные формулы, то возникает задача создания теста или системы тестов, которые позволили бы нам оценить ту или иную формулу в качестве "*хорошего*" индекса. Данная проблема очень остро обсуждалась в 20-е годы прошлого века. Тогда же были заложены *основы* современной теории *индексов*. В 1922 г. И. Фишер опубликовал известную статью **"***Построение индексов. Учение об их разновидностях, тестах и достоверности***".** Анализ проблемы *индексов*, проведенный Фишером, был настолько глубок, что практически все последующие исследования в данной области так или иначе опираются на эту работу.

Идея Фишера достаточно проста, а именно. В общем случае цены отдельных товаров за один период времени изменяются по-разному. Фишер заметил, что точки, соответствующие этим показателям (*отдельным индексам цен*), окажутся разбросанными, имеется явно выраженный *кластер* точек, отражающий среднее движение цен. Такой *кластер* обязательно существует, т.к. цены на отдельные товары не являются независимыми величинами. В ряде случаев эту взаимосвязь легко выявить, например между ценами на взаимозаменяемые товары, иногда ее можно выявить только на основе *специальных* статистических методов. Следовательно, в этом случае удобнее всего пользоваться средним показателем изменчивости, вместо изменения цен каждого отдельного товара.

Таким образом, ***индекс*** – это некоторая средняя из индивидуальных индексов цен. И. Фишер применил простой *исходный принцип*, который оказался очень эффективным. В соответствии с ним достаточно сложное явление, такое как, например, движение уровня цен, может быть изучено при помощи другого, которое подобно данному, но значительно проще его. Таким простым явлением в данном случае стало движение *индивидуальных* цен товаров. Существует много условий, которые с очевидностью выполняются для любого частного индекса. Фишер выбрал два из них в качестве формальных критериев (*тестов*) определения "*идеальности*" индексной формулы. Для нее тесты должны были выполняться без *систематической* ошибки.

*Первый* критерий И. Фишер назвал ***тестом обратимости по времени.*** Его суть заключается в том, что *произведение* прямого индекса цен на обратный ему должно равняться *единице*. Для индивидуальных индексов данное равенство, естественно, выполняется. Например, если в Таллине CD в два раза дороже, чем в Вильнюсе (*следовательно*, *в Вильнюсе* CD *в два раза дешевле*, *чем в Таллине*), то произведение этих двух индексов будет равно единице.

*Второй* критерий И. Фишера называется ***тестом обратимости факторов*** и формулируется как: индекс стоимости равен *произведению* индекса объема на индекс цен. Например, каждый понимает, что если он купил CD в два раза больше, чем год назад, но по вдвое меньшей цене, то количество затраченных им денег не изменилось.

*Первый* тест Фишера можно определить и так: *если в формуле* **PI** *поменять местами базисный и текущий периоды*, *то новый индекс должен равняться обратной величине исходного*. Этому тесту удовлетворяют *элементарные* индексы, но ему не удовлетворяют многие известные формулы агрегатных индексов, несмотря на его весьма простое логическое обоснование. *Второй* тест гласит: *произведение* **PI** и **VIT,** *построенных по одной и той же формуле*, *должно дать* **VIP.** Данный тест также имеет прозрачное логическое обоснование, однако он неприменим к агрегатным индексам, кроме индексов типа Пааше. Этим вопросам взаимосвязи индексов и показателей, принципам их построения и обоснования посвящено весьма много работ и данный вопрос выходит за рамки настоящей книги.

Фишер применил эти критерии к 134 индексным формулам, многие из которых он вывел впервые. Огромная вычислительная работа была им проделана на основе массива данных о *ценах* и *объемах* на рынках США по 36 товарным позициям за 6 предвоенных и послевоенных лет. Идеальной формулы И. Фишер не нашел, а именно: *не было ни одной средней*, *одновременно отвечающей предложенным тестам*. Это подтвердило его первоначальное предположение о том, что идеальной формулы *среднего индекса* не существует. Лучшей же оказалась формула, представляющая собой комбинацию индексов Ласпейреса и Пааше. Она получила название ***идеального индекса Фишера,*** рассмотренного нами выше.

Хотя идеальный индекс Фишера, ввиду этих логических критериев, казался бы наилучшим индексом, на практике он широко не используется. Мы упомянули уже одну из *причин* этого: индекс требует нового списка величин текущегох периода при каждом его вычислении. При этом, идеальный индекс не может дать тип сравнения, которое мы желаем выявить. Ценовой

индекс Ласпейреса полностью игнорирует изменение цен, потому что количества считаются постоянными. Индекс Фишера не делает этого, т.к. в его вычислении участвуют два набора величин и величины *текущего* периода изменяются для каждого нового вычисления индекса. В зарубежной статистике и экономике большинство важных индексов цен и величин, на сегодня, является различными разновидностями индекса Ласпейреса (75.2) или взвешенных средних ценовых отношений. Тогда как в отечественной экономике предпочтение отдается агрегатным индексам типа Паашее и взвешенным средним ценовых отношений. С учетом сказанного, рассмотрим вопросы взаимосвязи между экономическими индексами немного детальнее.

Сложность социально-экономических явлений, наличие многосторонних связей между ними приводит к тому, что отдельно взятый показатель не в состоянии сколько-нибудь полно отражать все богатство *содержания* этих явлений. Следовательно, отпадает идея *универсальных* показателей, ибо каждый отдельный показатель отражает только одну из сторон динамики явлений или одну из сторон *взаимосвязи* между ними. Только система индексов и показателей позволяет всесторонне отражать динамику явлений в их совокупности. В масштабе отрасли или всего народного хозяйства взаимосвязь индексов и показателей имеет особое значение, что требует разработки тщательно продуманной системы взаимосвязанных индексов. Речь здесь идет не о том, чтобы каждый индекс удовлетворял одним и тем же формальным требованиям, например, тестам И. Фишера, а о том, чтобы взаимосвязь *индексов* и *показателей* выражала реально существующие экономические взаимоотношения между явлениями и целыми отраслями. Указанная взаимосвязь позволяет измерять факторы, определяющие динамику того или иного явления. Так, объем произведенной продукции определяется как функция двух величин: производительности труда и затраченного рабочего времени. Таким образом, методы построения индексов продукции и индексов производительности труда должны взаимоувязывать оба указанных экономических индекса.

Алгебраически эта взаимоувязка характеризуется тем, что произведения двух индексов дают некий третий, т.е. в основе построения индексов лежит единый методологический принцип. Более того, этот принцип находит свое выражение в многообразии самих способов расчета и алгебраических форм индексов. Эти соображения лежат в основе отрицания отечественной статистикой увязывания индексов *формально-математическим* способом, не имеющим какого-либо удовлетворительного экономического обоснования.

Взаимосвязь между *цепными* и *базисными* **ИИ** уже обсуждалась выше и имеет как формальную основу, так и прозрачное экономическое обоснование. **АИ** качественных показателей (*цен*, *себестоимости*, *производительности труда и др.*) всегда являются индексами с переменными весами, ибо объем продукции каждый раз берется на уровне текущего периода. Поэтому к ним неприменим цепной метод расчета, а значит, и соответствующая взаимосвязь цепных и базисных индексов. Однако, в полном объеме эта взаимосвязь применима к **АИ** стоимости. Из простейших отметим две связи индексов. Между **PI, VIP** и **VIT** имеет место постоянное *соотношение* **VIP = PI*VIT,** которое получается, если **PI** вычисляется по формуле типа Пааше (**86.a, b**). Между **VIP, ILP** и индексом трудозатрат (**ILC**) существует подобное определяющее соотношение **VIP = ILP*ILC.** Например, из него следует, что увеличение трудозатрат в 1.25 раза, а **ILP** – в 2 раза, даст увеличение объема продукции в 2.5 раза.

Определение влияния каждого из *взаимоувязанных* факторов на совокупный экономический показатель составляет одну из сторон проблемы взаимосвязи индексов. Если все входящие в систему *индексы* и *показатели* выражены в одних и тех же единицах измерения, как это имеет место, например, при расчетах *производительности труда*, *задача* решается достаточно просто. Но в случае системы разнородных индексов и показателей анализ встречает существенные затруднения. Здесь задача статистики состоит в том, чтобы дать практикам статистикам и

экономистам наиболее простые методики расчета *влияния* отдельных факторов на конечный результат исследуемой ими производственной и/или экономической деятельности в той или иной области. К вопросу взаимосвязи индексов относится и задача определения влияния структуры совокупности на уровень индексируемой величины. Непосредственно данное влияние определить не представляется возможным. Поэтому, большинство статистиков и экономистов считают возможным делать это путем сопоставления результатов вычисления двух индексов с переменным и постоянным составом, о которых говорилось выше.

Проблема определения влияния *структуры* совокупности на *уровень индексируемой* величины также имеет совершенно бесспорное отношение к вопросу взаимосвязи индексов. Данное влияние невозможно для определения. Поэтому, большинство статистиков и экономистов считает возможным сделать это посредством *сравнения* результатов вычисления двух *индексов* – с *переменными* и *постоянными* структурами, о которых говорилось выше. Большое число работ посвящено вопросам взаимосвязи индексов и показателей, принципам их построения, обоснования и интерпретации; однако, данная проблема находится вне рамок настоящей книги. В частности, достаточно популярное обсуждение проблем построения, использования и интерпретации индексов цен может быть найдено в интересной книге [130].

*Индексы* относятся к важнейшим *обобщающим* показателям. С помощью *экономических индексов* можно исследовать динамику социально-экономического явления за два и более периодов времени, динамику среднего экономического показателя и сопоставлять уровни явления в пространстве: по странам, экономическим районам, областям и т.д. Индексы весьма широко используются также для определения *степени влияния измерений* значений одних показателей из фактических цен в сопоставимые. В статистической практике индексы наряду со *средними* величинами являются наиболее распространенными статистическими *показателями*. На базе индексных показателей решаются следующие основные задачи:

**1)** *характеристика общего изменения сложного экономического показателя* (*например*, *затрат на производство продукции*, *стоимости произведенной продукции и т.д.*) *или формирующих его отдельных показателей-факторов*;

**2)** *выделение в изменении сложного показателя влияния одного из факторов путем элиминирования влияния других факторов* (*например*, *увеличение выручки от реализации продукции*, *связанное с ростом цен или выпуска продукции в натуральном выражении*).

**3)** *в качестве самостоятельной выделяется задача обособления влияния изменения структуры явления на индексируемую величину*.

Способы построения индексов зависят от содержания изучаемых показателей, методологии расчета исходных статистических показателей, имеющихся в распоряжении исследователя статистических данных и самих целей исследования. Индексные показатели в статистике вычисляются на высшей ступени статистического обобщения и опираются на результаты сводки и обработки данных конкретного типа статистического наблюдения.

В заключении рассмотрения ***индексного метода,*** необходимо отметить, что при построении индексов часто возникают очень трудные проблемы определения, спецификации и выбора. Часто статистические методы затруднительны для использования при построении индексов. Более детально с индексным методом в статистике и экономике заинтересованный читатель может ознакомиться, например, в интересных книгах [49, 64, 72, 92, 116, 130, 157, 173, 182, 262, 270-272, 285].

# Глава 10.
# Компьютерные средства статистического анализа данных

Большинство статистических задач достаточно трудоемки и требуют большого количества вычислений, ограничивающихся в целом ряде случаев весьма простыми математическими операциями. Поэтому автоматизация решения такого класса задач имеет длинную историю, восходящую к простым приемам, облегчающим вычисления (*специальные статистические таблицы*, *приемы упрощения вычислений*, *механические*, *электрические и др.*), и развивающуюся совместно с прогрессом средств автоматизации умственного труда. Прежде всего, прогресс в этом направлении связан с развитием средств **ВТ** и связи, который в последние десятилетия получил мощный импульс под общим названием "*компьютеризация*" из-за массового использования обширного класса **ПК** [7-11, 13-22, 134-144, 190, 226, 227, 277].

Современное развитие указанных средств позволило решать общую задачу сбора, передачи, обработки и хранения различного рода данных, включая различного рода статистические данные, комплексно на основе локальных, ведомственных, региональных, национальных и транснациональных информационно-вычислительных сетей. В настоящее время сетевая технология обработки информации все шире используется в информационных системах различного назначения, в технике, в банковском деле, статистических органах и др. В этом плане можно отметить хорошо известную отечественным статистикам **АСГС** – *межотраслевую многоуровневую автоматизированную систему сбора*, *передачи и обработки учетно*-статистической информации, необходимой для планирования и управления народным хозяйством бывшего Союза. **АСГС** создавалась как одно из наиболее важнных функциональных звеньев **ОГАС** – общегосударственной автоматизированной системы. Организационно **АСГС** строилась по административно-территориальному принципу и включала союзный, республиканский, областной и районный уровни. В середине 80-х годов возникла объективная необходимость создания более эффективной *Единой статистической информационной системы* (**ЕСИС**), но в это дело весьма существенные коррективы внесла "*перестройка*".

Из обширного комплекса программно-технических средств поддержки сетевой обработки информации мы будем рассматривать только *программные средства* (**ПС**), допускающие, в первую очередь, *локальный* режим обработки информации, т.е. непосредственно на рабочем месте специалиста, решающего те или иные задачи статистики без использования средств *компьютерной телекоммуникации*. Тогда как с вопросами компьютерной телекоммуникации читатель может ознакомиться, например, в книгах [21, 136-138, 143, 226, 227]. В следующем разделе рассматриваются основные **ПС** и важнейшие подходы к решению разнообразных статистических задач.

## 10.1. Основные предпосылки использования компьютеров в статистике

Можно выделить шесть главных направлений применения компьютерной технологии для обеспечения задач анализа статистических данных, а именно:

| | |
|---|---|
| **A** | *хранение, обработка и верификация данных* |
| **B** | *сводка свойств данных* |
| **C** | *анализ данных средствами компьютерной технологии* |
| **D** | *вычисления в процессе использования статистических процедур* |
| **F** | *анализ данных на основе модельного подхода* |
| **E** | *сетевой режим обработки статистических данных* |

**A. *Хранение, обработка и верификация данных*.** При многократной обработке больших объемов данных весьма актуальной является проблема их *хранения*. Способ хранения данных определяет многие очень важные свойства работы с данными, а именно: *надежность хранения*, *скорость доступа к данным*, *эффективность обновления данных*, *защита от несанкционированного доступа*, *дублирование данных* и т.д. Если данные находятся в форме отчетов, собранных за длительный период времени, или если используются большие объемы данных, необходимо решить следующие важные вопросы: (**1**) ввод данных и механизм их хранения должны быть достаточно гибкими, чтобы обеспечить при необходимости эффективное их обновление; и (**2**) *программа*, обеспечивающая ввод и обработку данных должна быть *мобильна* относительно вычислительной платформы (*Windows*, *Linux* и др.). Наряду с этим, программа ввода данных должна быть снабжена достаточно развитой системой проверки (*верификации*) данных.

Прежде всего, необходимо тщательно рассмотреть *характер* и *объем сохраняемой* информации. К решению проблемы хранения данных во внешней памяти и организации доступа к ней целесообразно проконсультироваться со специалистом по *системам управления базами данных* (**СУБД**). Данная область прикладной информатики довольно хорошо развита и обеспечена большим количеством **СУБД** различного уровня и назначения. Как правило, статистические агентства используют ту или иную **СУБД** и, следовательно, в большинстве случаев обычный прикладной статистик уже *ограничен* в этом направлении выбранной технологией хранения данных и доступа к ним.

В качестве типичного примера современной *системы управления базами данных* (**СУБД**) можно представить известную система ***Oracle*** фирмы *Oracle Corp*. [242]. ***Oracle*** является объектно-реляционной **СУБД**, которая обеспечивает открытый, всесторонний, и интегрированный подход к информационному управлению. ***Oracle*** состоит из непосредственно *базы данных* (**БД**) ***Oracle*** и сервера ***Oracle***. *Сервер* базы данных **–** *ключ* к решению задач информационного управления. В целом, сервер должен надежно управлять большим количеством данных в многопользовательской среде так, чтобы много пользователей могли одновременно получать доступ к тем же самым данным. Все это должно быть достигнуто при обеспечении высокой скорости обслуживания запросов к данным. Сервер должен запрещать *несанкционированный* доступ и обеспечивать эффективную обработку отказов. Сервер ***Oracle*** эффективно решает указанные задачи хранения и обработки данных произвольной организации.

Уже только на основе перечисления возможностей ***Oracle*** [242] можно с весьма определенной уверенностью говорить о целесообразности использования **DBMS**-*технологии* для *организации* хранения данных и доступа к ним, включая данные, предназначенные для статистического анализа. Использование **DBMS**-технологии устраняет много проблем хранения данных и доступа к ним, и обеспечивает эффективные процедуры ряда других важных компонентов работы с данными различной структуры и характера, включая поддержку работы с данными в сетевом режиме. Однако, более детально данный вопрос здесь не рассматривается. Поэтому заинтересованный читатель отсылается к руководствам по соответствующим статистическим пакетах или, например, к книгам [18, 20, 81, 138, 142, 190, 239-241]. Более того, в реферативных материалах [227] читатель регулярно может получить информацию по новым книгам по вопросам хранения данных и доступа к ним.

**B.** ***Сводка свойств данных.*** Увеличивающееся использование таблиц, диаграмм различных видов и сводки свойств данных для описания *особенностей* наборов данных вызвало быстрый рост числа статистических пакетов для поддержки необходимых процедур. Многие из этих статистических средств будут кратко рассматриваться ниже. Для пользователя можно дать только одно предупреждение, а именно: тщательно выбирать *метод* или *способ сводки данных*. Успешный выбор метода *сводки данных* позволит более успешно проводить их последующий статистический анализ, включая процесс самого *планирования* анализа. Простота применения простых статистических пакетов соблазняет к желанию получить массу *итоговой информации*, относящейся к анализируемым данным. Между тем, необходимо иметь в виду, что слишком много информации может скрыть важнейшие особенности данных, также как и недостаток такой информации.

**C.** ***Анализ данных средствами компьютерной технологии.*** Анализ данных является разделом общего статистического анализа, который развивался наряду с компьютерной технологией. Разумное использование компьютерной технологии позволяет помочь с ответом на многие экспериментальные вопросы и предложить *новые вопросы*, которые могут вести к результатам, не предсказанным исследователями. Опасность процесса состоит в сложности распознавания *реальных* эффектов от *выборочных данных*. Практически любой набор данных, исследованный со многих сторон, выявит модель, которая могла бы иметь научное значение.

Необходимо отметить, что с практической точки зрения имело место медленное развитие стандартных пакетов, которые предлагают формализованные методы исследовательского анализа данных. Более высокий уровень анализа данных возникает, когда предлагаются дедуктивные статистические модели, а анализ используется для исследования приемлемости предположений в модели, которые являются *определяющими* для правильного использования статистических методов.

Все шире этот вид анализа данных используется в области, которая может быть определена как подгонка моделей. Например, в регрессионном анализе и изучении отношений между переменными классический статистический анализ данных связан с выбором подходящей модели из определенного набора возможных моделей (*в частности*, *линейная или нелинейная модель*). Один из возможных подходов к такой компьютерной технологии подгонки моделей *регрессии* и определения *трендов* временных рядов рассматриваются в разделах **7.5** и **8.4** книги на основе пакета *Maple*. Элементы данной компьютерной технологии исследовательского анализа данных будут рассматриваться ниже при обсуждении основных функциональных статистических средств пакета *Maple*.

**D.** ***Вычисления в процессе использования статистических процедур.*** За *редким* исключением, обычно используемая статистическая техника и процедуры очень удобны для их реализации средствами компьютерной технологии. Болеее того, их *большая* часть не требует применения очень сложных вычислительных методов и может быть запрограммирована пользователем с довольно низким уровнем опыта в программировании. Учитывая, что для большого числа массовых статистических процедур, например, вычисления средних, большинство систем программирования и пакетов прикладных программ имеют встроенные функции или процедуры. В свою очередь, некоторые из статистических методов были бы невозможны без использования соответствующих средств программного обеспечения.

**E.** ***Анализ данных на основе модельного подхода.*** Знание свойств статистических методов восходит, в основном, к математическому анализу. Однако, между некоторыми важными статистическими методами и их математическим обоснованием имеется взаимозависимость. В таких случаях достаточно важную роль в обеспечении решений математических задач могут играть компьютеры, неоднократно моделируя данные модели и исследуя свойства

полученных наборов данных. Использование компьютеров играет особую роль при сборе и статистическом анализе данных, полученных в результате *любого* эксперимента, проводимого на оборудовании, связанном с компьютерами специальными датчикамив. В таком случае имеется возможность не только *проведения* соответствующего статистического *анализа данных*, полученных в результате эксперимента, и оценки их соответствия эмпирической модели исследуемого явления, но также и реализации обратной связи, обеспечивающей диалоговый контроль корректности (*чистоты*) проведения эксперимента.

**F. *Сетевой режим обработки статистических данных.*** Данный режим играет все большую роль в статистическом анализе; прежде всего, в статистических организациях, соединенных электронной сетью передачи данных как друг с другом, так и с источниками статистической информации более низкого уровня. Деятельность современных органов статистической службы в значительной мере основана на *сетевом режиме* обработки статистических данных. Для успешного приенения компьютерной технологии требуется выполнение ряда условий, с которыми детальнее можно ознакомиться, например, в [347] и приведенных в ней ссылках.

Важная компонента любого компьютерного статистического анализа – выбор необходимых для его выполнения программного обеспечения. Вообще говоря, есть три группы программ, ориентируемых на статистический анализ данных, а именно:

| | |
|---|---|
| **1** | *статистические пакеты, которые содержат основные формы ввода данных, средства обеспечения наиболее всестороннего статистического анализа данных и вывода результатов в требуемых форматах* |
| **2** | *специализированные статистические пакеты (программы), которые предназначены для специального статистического анализа данных* |
| **3** | *программы, созданные пользователем для решения собственных статистических задач* |

Средства *первой* группы носят достаточно универсальный характер и ориентированы на довольно широкий круг статистических приложений. Однако, их освоение требует, порой, *немалой* квалификации и, как правило, существенных временных затрат. В качестве примера такого типа можно привести известный статистический пакет **SAS**. Средства *второй* группы не носят универсального характера и ориентированы на довольно узкий круг приложений. Освоение средств второй группы, как правило, не требует таких существенных усилий, как это имеет место для средств *первой* группы. В качестве примера средств *второй* группы можно привести пакет **COMFORT,** являющийся диалоговой системой прогноза, ориентированной на решение проблемы подготовки прогнозов специалистами по прогнозированию. Однако, использование специализированных статистических средств может дать в результате ряд нежелательных последствий таких как: несовместимость структур данных и вычислительных платформ, существенные несовместимости языков управления и т.д., которые могут вызвать довольно существенные затруднения при их использования в сети расширенного анализа статистическихо данных. Потенциальные неудобства специализированных пакетов должны быть ясно поняты и осмыслены до принятя решения об их использовании в качестве средств статистического анализа данных.

Наконец, программы, созданные пользователями для собственных статистических нужд, составляют средства *третьей* группы. Средства данной группы в значительной степени несут индивидуальный характер и имеют достаточно узкий круг приложений, хотя они и могут довольно эффективно решать специфические статистические задачи в той или иной области приложений. Ряд такого вида средств рассматривается в книге в форме иллюстративных статистических процедур, реализованных в среде пакета *Maple*. Между тем, такие программы пользователя могут существенно дополнять и универсальное, и специальное статистическое программное обеспечение, повышая эффективность статистического анализа в целом.

Все чаще использование статистических пакетов различных уровней для анализа данных становится весьма существенной частью статистического анализа данных в целом. По мере их доступности и пригодности для использования на **ПК,** исследователи, использующие статистический анализ данных, имеют возможность делать собственный выбор пакета, если более, однако, руководящие инстанции (*статистический отдел*, *центральный вычислительный центр и т.д*) предварительно не предусмотрели этого. В этой связи, необходимо помнить, что основные характеристики выбранного пакета должны быть адекватны потребностям и опыту пользователя. Обоснование критериев такого выбора – достаточно многоаспектная проблема и ее рассмотрение не входит в планы настоящей книги. Заинтересованный же читатель отсылается за информацию к книгам [20, 44, 47, 81, 126, 132, 138, 142]. В следующем разделе кратко рассматривается *базовое* статистическое программное обеспечение и подходы к решению различных статистических задач.

## 10.2. Краткий обзор статистического программного обеспечения

Средства для решения статистических проблем можно классифицировать (*с определенной степенью условности*) на четыре основных группы, а именно:

| | |
|---|---|
| **1** | *библиотеки стандартных статистических программ* (**LSSP**) |
| **2** | *встроенные средства систем программирования и пакетов* |
| **3** | *универсальные пакеты прикладных программ* |
| **4** | *специальные пакеты прикладных программ*, *включающие специальные языки для программирования статистических задач* |

К *первой* группе средств относится большое количество программ статистического анализа, реализованных на различных языках программирования для различных классов и типов компьютеров. Как правило, **ПС** данной группы требуют определенной программистской культуры и используются как в автономном режиме, так и в виде программ в программных системах пользователя. Типичным представителем средств данной группы является библиотека **ПНП-БИМ** института математики АН СССР, работавшая на ЕС ЭВМ и М-4030. Библиотека содержит более 1000 стандартных подпрограмм по численным методам и математической статистике. В настоящее время расширеннаяых версия данной библиотеки реализована и для **ПК.** К средствам данной группы можно также отнести библиотеки **SSP** и **NAG,** написанных на Фортране и реализованных для всех классов компьютеров. Библиотеки содержат много подпрограмм числового анализа, включая обширные средства обеспечения статистического анализа данных. Однако при использовании средств данной группы нужна определенная программистская культура, в общем случае не присущая *непрофессиональному* в этом отношении пользователю, например, статистику, экономисту, финансисту и др.

Средства *второй* группы характеризуются как встроенные функции того или иного вида процедур статистического анализа в системах программирования или пакетах. В качестве примеров **ПС,** содержащих средства второй группы, можно привести известные системы программирования ***Turbo-Pascal*** [19, 138]**,** ***Turbo-Basic*** [11, 13, 17]**,** известные математические пакеты ***MathCAD*** [11, 14, 15, 17, 228]**,** ***MatLAB*** [229, 233, 235]**,** ***Mathematica*** [99, 134-138]**,** ***Maple*** [62, 97, 98, 139-144, 158, 230, 233, 286-290, 302], ***Macsyma*** [124, 234] и др.

Перед дальнейшим изложением сделаем одно небольшое пояснение. Как правило, в общем случае решение статистической задачи включает следующие базовые процедуры: подготовка и редактирование исходных данных (*быть может с использованием* **СУБД**), собственно сам статистический анализ, представление результатов в *обычном*, *графическом* и/или *табличном* виде. Исходя из этого, *пакеты прикладных программ* (**ППП**), обеспечивающие *все* виды

указанных процедур, будем называть *интегрированными* или *универсальными*, в противном случае *специальными* или *узко-специальными*, если они ориентированы, в основном, на один тип процедур, например, для графического анализа данных.

Наконец, средства *третьей* группы составляют универсальные **ППП** для статистического анализа данных. Из наиболее известных можно отметить следующие пакеты: **MicroStat, Soritec, Statistica, STATA, Systat, StatGraf, P-Stat, StatPro, StatView, MultiStat, SAS, DataStat** и др. Данные пакеты имеют различные структуру и интерфейс с пользователем, обеспечивая весьма широкий набор статистическиих процедур для анализа данных, наряду с другими функциями (*ввод и редактирование данных*, *графический анализ и др.*). К пакетам, как правило, прилагаются различные библиотеки программы, позволяющие управлять статистическими процедурами пользовательских программ; однако, это уже требует определенных навыков.

При этом, некоторые из пакетов имеют *встроенный* язык программирования, ориентируемый на *реализацию* некоторого вычислительного алгоритма статистического характера. Хорошим примером может служить диалоговая статистическая вычислительная среда пакета **XploRe** [169]. **XploRe** представляет сочетание как классических, так и современных статистических процедур в совокупности со сложной интерактивной графикой. Пакет **XploRe** представляет хорошуюй основу для статистического анализа данных, исследования и обучения. Его цель состоит в исследовании и анализе данных, так же как в разработке новых статистических методов. Кроме того, пакет включает объектно-ориентированный язык программирования высокого уровня; пользователь может включать в пакет собственные методы.

В заключении представления средств *третьей* группы могут быть упомянуты два наиболее интересных статистических средства. Достаточно широко используемый пакет ***Microsoft Excel*** включает набор процедур анализа данных (*названный* ***модулем анализа***), служащий для решения достаточно сложных статистических и технических задач. Для реализации анализа данных посредством процедур данного модуля необходимо указать анализируемые данные, представленные рядами или графами ***Excel*** -листа, и выбрать нужную функцию анализа данных. Однако, необходимо иметь в виду, что требуется начальное статистическое знание для успешного использования этих процедур анализа. В качестве инновационного развития вышеупомянутого модуля анализа служит пакет **S-Plus** фирмы Mathsoft Inc [45, 282]. **S-Plus** успешно сочетает использование стандарта *Microsoft Office*-совместимых программ с *методами* визуализации и широким набором классических и робастых статистических функций.

Среди средств *третьей* группы одним из бесспорных лидеров является известный пакет **SAS** (*Statistical Analysis System*) [34, 38, 100, 101], изначально созданный в Университете Северной Каролины (США) в 1979 г. для статистической обработки данных на универсальных ЭВМ. Со временем пакет был преобразован в довольно гибкую и широко функциональную систему, предназначенную для пользователей с ограниченными статистическим и компьютерным опытом. Пакет **SAS** постоянно развивается и в настоящее время включает 22 подсистемы, ориентированные на различные приложения [34, 38, 100, 101, 242]. Согласно возможностям, функциям и архитектуре пакет **SAS** представляет интегрированную систему для поддерки фундаментального статистического анализа различного вида данных. Пакет поддерживает работу почти со всеми классами и типами компьютеров и автоматизированных рабочих мест, с различными операционными платформами, имеет развитый многопользовательский интерфейс и поддерживает мощную среду для разработки приложений, удовлетворяющих, практически, любого пользователя в любых разделах бизнеса, экономики, финансов и др. В настоящее время пакет поддерживается SAS Institute Inc (США) [34, 38, 100, 101]. Пакет **SAS** используется Департаментом Статистики Эстонии совместно с системой управления базами данных **ADABAS** как локальном, так и в сетевом режиме статистической обработки данных и решения различных задач государственной статистики.

В отличие от западных, многие отечественные пакеты в гораздо большей степени подходят для нужд среднего российского пользователя. Здесь основные операции, как правило, сразу обозримы из головных меню, а рутинные процедуры выполняются с минимумом действий и разветвлений по принципу: «прямым путем – к понятному результату». Наиболее развитой системой контекстной экранной помощи, включающей объемный справочник-гипертекст и экспертную систему по выбору метода статистического анализа, обладает пакет **STADIA**. Здесь каждый числовой статистический вывод сопровождается короткой и весьма понятной интерпретацией (*более искушенный в статистике пользователь может сделать интерпретацию результатов самостоятельно, ибо все данные для этого выводятся на экран*). В пакете **Мезозавр** реализована оригинальная система экспертной оценки сложных моделей временных рядов. Система **Эвриста** выделяется живо и изобретательно написанной документацией, которая читается как захватывающее повествование о возможностях статистических методов. Все три пакета аккумулируют передовой опыт российской науки, что не удивительно: их создавали ведущие специалисты Академии наук и МГУ. Они эксплуатируются сотнями пользователей на протяжении целого ряда последних лет. Периодически на выставках и семинарах можно встретить и многие другие интересные российские программы анализа данных. Можно еще отметить целый ряд интересных отечественных пакетов таких как: **COPPA, SOD-GS, PAD, SIAD, OTEKS, DIAS, TIPOLOG, ELIMINATION, STP, STADIA, STATAN,** которые по целому ряду характеристик существенно превосходят зарубежные аналоги.

Наконец, *четвертая* группа средств включает пакеты, ориентируемые на решение отдельных классов статистических задач и/или разделов общей технологии статистического анализа данных в целом. Так, пакет прикладных программ **BMDP** (*Биомедицинские Программы*) [50], первоначально созданный в Университете Калифорнии (США) в 60-х, был ориентирован на статистический анализ *биомедицинских* данных; однако, его последующее развитие позволило использовать пакет в качестве достаточно общего средства для реализации различного вида статистического анализа в других прикладных областях. Последние версии пакета включают средства, обеспечивающие чрезвычайно широкий набор элементарных и довольно сложных статистических процедур. При этом, *язык управления* пакетом **BMDP** ориентирован на такого пользователя, который не является программистом. К пакету **BMDP** по целям и функциям примыкает и японский пакет **SPMS**, в котором реализована *специализированная* база данных.

Пакеты **SPSS** и **BMDP,** ориентируемый первоначально на статистический анализ данных в специальных прикладных областях, в течение собственного развития, практически, перешли на третий уровень универсальных статистических **ППП**. На сегодня они составляют группу наиболее используемых статистических **ППП,** а именно: **SAS, SPSS** и **BMDP**. Превосходный сравнительный анализ этих трех статистических пакетов на примерах решения различных задач анализов: регрессионного, дисперсионного, корреляционного, дискриминантного и группового может быть найден в вышеупомянутой книге [35].

Пакет **ITSM** является наиболее высоко специализированным пакетом. Он предназначен для анализа временных рядов, представляя достаточно большой интерес для анализа данных в целом ряде дисциплин таких как: статистика, математика, бизнес, биологиия, медицина, техника и другие естественные и социальные науки. Пакет **ITSM** позволяет решать задачи моделирования временных рядов и задачи прогнозирования [36].

Три *полезных* пакета ***REGRESSION*, *DESCRIPT* Statistics** и ***BUSINESS* Statistics** предназначены для задач регрессионного анализа, *описательной* и *экономической* статистики соответственно. ***ECONOMIST* Workstation** является *автоматизированным рабочим местом* (**АРМ**) экономиста. В СССР был создан **АРМ *Statistician*,** ориентированный как на статистическую обработку, так и на *анализ* экономической информации. Основными *функциями* этого интересного **АРМ**

являются: создание и обработка различных типов, графический анализ, математическая и статистическая обработка различных видов данных.

Для графического представления результатов статистического анализа данных, разработано много интересных пакетов Таких как: **GR-7** – графический анализ временных рядов, **CBIG** – интегрированная графическая система, **SABL** и **IMGRAF** – интерактивная графическая система анализа данных, **Diagram Microsoft Graph** и целый ряд других. Учитывая важность графического представления статистических данных, **ППП** данной группы представляются нам довольно важными средствами при решении задач статистического анализа данных в различных прикладных областях.

В ходе дальнейшего развития статистического анализа данных потребовались средства для преобразования файлов данных в форматы других статистических пакетов, обеспечивая их информационную совместимость, а также для преобразование входных данных сложной структуры (*например*, *расположенных в базах данных*) в форматы известных статистических пакетов. Известными пакетами такого типа являются **SIR, ACIS, SLANG, LEDA, WRAPS** и др. Пакеты **RGSP, STDERR, SUPERCARP, SEPP, CLUSTER** и ряд других поддерживают автоматизацию статистической обработки данных сложных выборочных наблюдений, в то время как пакет **EISPACK** занимает довольно специальное место и поддерживает числовые методы статистического анализа данных. Для решения задач статистического анализа могут успешно использоваться также пакеты другого типа. Так, например, известный пакет **GPSS** используется для моделирования сложных статистических систем различного типа [185].

В заключение настоящего раздела кратко охарактеризуем наиболее используемые в России статистические пакеты ***STATISTICA*** и ***STADIA***. ***STATISTICA*** – универсальная *интегрированная* система, предназначенная для статистического анализа, визуализации данных и разработки пользовательских приложений. Программа содержит широкий набор процедур анализа для применения в научных исследованиях, технике, бизнесе. Помимо общих статистических и графических средств в системе имеются специализированные модули, например, для проведения социологических или биомедицинских исследований, решения технических и промышленных задач, а именно: анализ процессов и планирование эксперимента, карты контроля качества. Пользователями системы являются крупнейшие университеты, научные центры, компании, банки всего мира, государственные учреждения. **ППП** ***STATISTICA*** может служить не только эффективным инструментом для научных исследований, но и чрезвычайно удобной средой для обучения методам статистического анализа. ***STATISTICA*** активно используется в учебном процессе в таких вузах, как МГУ, МГИЭМ, МЭСИ, МФТИ, МИФИ, МГТУ им. Баумана, СпбГУ ЭФ и многих других. ***STATISTICA*** является наиболее динамично развивающимся статистическим **ППП** и по многочисленным рейтингам является мировым *лидером* на рынке статистического программного обеспечения. Пользователь может добавить *собственную панель инструментов* с тем или иным *методом* статистического анализа. Несомненным достоинством пакета является возможность расширять систему при помощи встроенного языка программирования. **STATISTICA Neural Networks** – универсальный и мощный *нейронно-сетевой* пакет. Он дает возможность автоматически получать эффективные и правильные решения для широкого круга задач, в которых использование традиционных статистических методов затруднено, например, из-за отсутствия априорных моделей либо конкретных гипотез.

В завершении отметим, что в рамках программы Экономического Комитета ООН *европейской* группой программного обеспечения был создан пакет **StatWare** представляющий собой специализированную системы управления базами данных, база данных которой содержит обширную информацию по программноmуо обеспечениюя, которое включает функции,

интересные для статистического анализа данных. Руководство и работа с пакетом описаны в нашей книге [20]. Там же можно найти полезную информацию по расширенному списку наиболее используемых статистических пакетов прикладных программ. Дополнительная полезная информация по статистическому программному обеспечению и математическим методам, используемым при их создании, может быть найдена, например, в книгах [44, 47, 56, 81, 128, 132, 156, 166, 190, 225, 236-238, 277].

В период существенного увеличения реальной роли принятия решения, прогноза и оценок результатов экономической, финансовой и иной хозяйственной деятельности производства, особая роль принадлежит статистическому анализу данных различного уровня и назначения с использованием различных типов и классов *компьютеров* и, прежде всего, обширного парка **ПК,** находящихся в институтах, различных организациях, офисах, фирмах, отделах и др. Много функций статистической обработки и анализа включены и в **ППП** соответствующего статистического характера, и в средства программного обеспечения другой цели. Изобилие *специальных* программных средств для решения статистических *задач* различного назначения, и подходящих статистических функций в программном обеспечении ставит вопрос поиска и выбора наиболее приемлемых статистических средств для конкретных условий применения (*текущие и перспективные задачи*, *технические средства*, *состав пользователей и т.д.*).

## 10.3. Использование класса персональных компьютеров в статистическом анализе

Программное обеспечение для статистического анализа данных, упомянутое в предыдущем разделе, в целом, *первоначально* было реализовано для *универсальных* ЭВМ (*UNIVAC*, *Borrough*, *IBM* 360/370 и т.д.); затем значительная часть *наиболее* популярных из них была *адаптирована* к *миникомпьютерам* (*PDP*-11, *WANG* и т.д.). Наконец, с появлением класса **ПК** статистическое программное обеспечение было *адаптировано* также и к этому классу *массовых* компьютеров, в *дополнение* к вновь созданным пакетам типа ***Systat***, ***StatGraf***, ***StatPro***, ***STADIA***, ***MultiStat*** и т.д. Появление класса **ПК** позволило (*насколько это вообще возможно*) приблизить статистическое программное обеспечение к непосредственному пользователю (*статистик*, *экономист*, *клерк*, *финансист и т.д.*), требуя *взамен* просто ознакомления с *основами* компьютерной грамотности. Это послужило серьезной предпосылкой повышения производственной эффективности не только больших фирм и корпораций, а также и значительно меньших фирм и организаций, обладающих **ПК** различных типов.

Используя соответствующее программное обеспечение, пользователь **ПК** имеет возможность не только быстро и качественно подготовить необходимые статистические отчеты и расчеты в соответствующие офисы и отделы (*отдел статистики*, *налоговый отдел*, *больничный фонд*, *отдел социального обеспечения и др.*), но также и решать задачи более *глубокого* статистического анализа финансовой, экономической и хозяйственной деятельности производства, включая ее прогнозирование. В то время как использование **ПК** в сетевом режиме позволяет работать со статистическими, финансовыми и банковскими системами в интерактивном режиме.

Отдельно следует сказать по поводу одного аспекта использования **ПК** для статистического анализа. Максимальная близость **ПК** к пользователю, который не является профессионалом в программировании, позволяет не только быстрее и эффективнее справляться с задачей приобретения компьютерной грамотности, но также и работа с некоторым статистическим **ППП** позволяет более глубоко освоить методы статистического анализа данных, включая самые сложные статистические процедуры. В то же самое время, такая работа великолепно стимулирует дальнейший рост профессионализма пользователя как в статистическом, так и в компьютерном отношениях.

Наконец, персональный характер использования компьютеров стимулировал дальнейшее развитие периферийных устройств **ПК** и, прежде всего, качество цветного видео-монитора, основной и внешней памяти, и средств входа/вывода, включая средства поддержки сетевого режима. Это, в свою очередь, позволило создавать высококачественные *графические* системы, высококачественную цветную печать и быстродействующие базы данных большого объема. В этой связи, *интегрированные пакеты* прикладных программ статистического анализа данных наряду с различными статистическими процедурами позволяют достаточно эффективно выполнять графический анализ данных (*чрезвычайно наглядный для пользователя*), выпускать высококачественные печатные материалы, использовать мощные базы данных различного назначения и работать как в локальном режиме, так и в информационных сетях различных уровней и организаций таких как: локальных, ведомственных, областных, региональных, национальных и транснациональных.

Возможности современных **ПК** наряду с их относительно низкой стоимостью позволяют создавать автоматизированные рабочие места статистика, экономиста и финансиста уже на самых нижних уровнях статистического и экономического анализа, обеспечивая создание многоуровневых автоматизированных систем (*рабочее место*, *предприятие*, *отрасль*, *область и т.д.*) обработки и анализа финансовой-экономической и статистической информации. Это замечание может быть отнесено, в первую очередь, к системе Государственной статистики, к банковским системам, налоговой системе, здравоохранению и ряду других.

Выбор **ПК** определяется характером его последующего использования – типами и классами задач, предполагаемых к решению. На основе такого анализа выбирается *наиболее адекватное* и прикладное, и системное программное обеспечение, и аппаратные средства ЭВМ с учетом определенной перспективы. Между тем, ввиду очень быстрого прогресса как программного обеспечения, так и аппаратных средств ЭВМ, необходимо отдавать предпочтение клонам наиболее перспективных и стабильных типов **ПК**. В результате такого анализа должна быть выбрана одна из компонентов программного обеспечения, которой ориентировалось бы на статистический анализ данных. В качестве подобных средств можно предложить **ППП SAS, StatGraf, STATISTICA, STADIA, Microsoft Excel, S-Plus** и др. Выбор, как правило, определен компьютерной грамотностью и возможностями пользователя, и областью его статистических интересов и приложений.

В любом случае, после *выбора* необходимого статистического программного обеспечения вам потребуютсяего изучение и определенная практическая наработка с ним, как предпосылка его дальнейшего успешного использования. Опыт показывает, что для хорошего освоения сложных статистических **ППП** требуется несколько лет наработки с ними. Естественно, в рамкахнашего курса "*Общая Теория Статистики*" не представляется возможным провести приличное знакомство даже с одним из средств подобного типа, за исключением их общего обзора. Однако, с методической точки зрения было бы довольно целесообразно получить некоторое понятие об использовании ПК для автоматизации решения задач статистического анализа данных. Именно поэтому мы избрали несколько нетрадиционный подход.

Учитывая недостаточную компьютерную грамотность, прежде всего, студентов социальных наук, мы создали набор простых *Maple*-документов и процедур, позволяющих выполнять элементарный статистический анализ в среде вышеупомянутого пакета *Maple* [97-99, 139-141, 143, 144, 158, 230, 233, 262]. Ранее, подобная работа была выполнена и для известного пакета *MathCAD* [14, 15, 17, 18, 29, 127, 228]. Вообще говоря, данные документы не ориентированы на коммерческое использование, и предназначены для иллюстрации использования **ПК** для статистического анализа данных в *объеме материала* настоящей книги. Однако, в ряде случаев они могут служить некоторыми образцами или основой для автоматизации решения задач

статистического анализа данных для читателя, имеющего доступ к пакету *Maple*, который является весьма широко распространенным в университетской среде. В следующих разделах рассматриваются элементы статистического анализа данных в среде вышеуказанного пакета *Maple*, включая его краткую характеристику. Современные теоретические и прикладные проблемы вычислительной статистики получили достаточно полное отражение в трудах международной конференции *COMPSTAT*-2000 [298].

## 10.4. Краткая характеристика математического пакета *Maple*

В настоящее время *программное обеспечение*, предназначенное для решения *математических* задач (*под* ***математической*** *задачей понимается любая задача, если алгоритм ее решения может быть описан в терминах любого раздела математики*), является довольно обширным и может быть условно дифференцировано на **5** уровней, а именно: (**1**) *встроенные средства различного уровня некоторой системы программирования;* (**2**) *специальные языки программирования;* (**3**) *узкоспециальные*, (**4**) *специальные* и (**5**) *универсальные математические пакеты*. Достаточно детальный обзор средств каждого из перечисленных уровней дается в книгах [136-145, 190]. Здесь же мы акцентируем внимание на средствах *пятой* группы, обеспечивающих решение различных задач продвинутой математики.

Среди современного программного обеспечения *пятого* уровня, ориентируемого на решение задач математического характера как в *числовом*, так и в *алгебраическом* виде, можно выбрать группу пакетов, являющихся наиболее развитыми и поддерживающими самые популярные платформы, а именно пакеты: ***MathCAD, Macsyma, Mathematica, REDUCE, Maple, MuPAD, Derive*** *и* ***AXIOM***. Детальнее с пакетами этой группы читатель может ознакомиться на основе соответствующих сайтов Интернета, тогда как наиболее развернутый сравнительный анализ ведущих систем *компьютерной алгебры* (***Maple, Macsyma, Reduce, Derive, Axiom, Mathematica***) может быть найден в вебсайтах [303, 304]. В то же время, *три* пакета ***Maple, Mathematica*** и ***MathCAD*** в последние годы получают все большее распространение как в исследованиях, так и в прикладных и образовательных целях. Между тем, наш опыт апробации и использования пакетов ***MathCAD, REDUCE, Maple*** и ***Mathematica*** позволяет говорить о пакетах ***Mathematica*** и ***Maple*** как о бесспорных лидерах (*на основе обобщенного индекса*) среди всех вышеназванных средств систем компьютерной алгебры [302].

*Maple* способен решать большое число, прежде всего, *математически ориентированных* задач вообще без программирования в общепринятом смысле. Вполне можно ограничиться лишь описанием алгоритма решения своей задачи, разбитого на отдельные последовательные этапы, для которые *Maple* имеет уже готовые решения. При этом, *Maple* располагает большим набором процедур и функций, *непосредственно* решающих вовсе не тривиальные задачи как-то интегрирование, дифференциальные уравнения и др. О многочисленных приложениях *Maple* в виде т.н. *пакетов* и говорить не приходится. Тем не менее, это вовсе не означает, что *Maple* не предполагает программирования. Имея собственный достаточно развитый язык программирования, пакет позволяет программировать в своей среде самые разнообразные задачи из различных приложений.

Входной язык ориентирован на решение *математически ориентированных* задач практически любой сложности в интерактивном режиме. Он обеспечивает диалог пользователя со своей вычислительной компонентой (*вычислителем*), принимая запросы пользователя на обработку данных с их последующей обработкой и возвратом результатов в символьном, числовом или графическом видах. *Входной* язык является языком *интерпретирующего* типа и идеологически подобен языкам этого типа. Он располагает большим числом математических и графических

функций, и другими средствами из обширных библиотек пакета. Интерактивный характер языка позволяет легко реализовать с его помощью интуитивный принцип решения своих задач, при котором ход решения можно пошагово верифицировать, получая в конце концов требуемое решение. Уже введя первые предложения в текущий сеанс пакета, вы начинаете работать со *входным Maple*-языком. Все примеры применения представленных в книге *средств* (*процедур* и *модулей*) являются типичными предложениями входного *Maple*-языка.

Среда программирования пакета обеспечивается встроенным *Maple*-языком, являющимся функционально полным процедурным языком программирования четвертого поколения (**4GL**). Он ориентирован, прежде всего, на эффективную реализацию как системных, так и задач пользователя из различных математически-ориентированных областей, расширение сферы приложений пакета, создание библиотек программных средств и т.д. Синтаксис языка наследует многие черты таких известных языков программирования как ***C***, ***FORTRAN***, ***BASIC*** и ***Pascal***. Поэтому пользователю, в той или иной мере знакомому как с этими языками, так и с программированием вообще, не составит особого труда освоить *Maple*-язык пакета.

Средства *Maple*-языка позволяют пользователю работать в среде пакета в двух режимах: (**1**) на основе функциональных средств языка с использованием правил оформления и работы с *Maple*-документом предоставляется возможность на интерактивном уровне формировать и выполнять требуемый алгоритм вашей задачи без сколько-нибудь серьезного знания даже основ программирования, а подобно конструктору собирать из готовых функциональных компонент входного языка на основе его синтаксиса требуемый вам алгоритм, включая его выполнение, отображение результатов на экране (в обычном и/или графическом видах), в файле и в твердой копии, и (**2**) использовать всю мощь языка как для создания развитых систем конкретного назначения, так и средств, расширяющих собственно саму среду *Maple*, чьи возможности определяются только вашими собственными умением и навыками. При этом, первоначальное освоение *Maple*-языка не предполагает предварительного серьезного знакомства с основами программирования, хотя знание их и весьма предпочтительно.

*Maple*-язык включает большое число математически ориентированных функций, позволяя только одним вызовом функции решать сложные самостоятельные задачи, а именно: решать системы дифференциальных уравнений или алгебраических уравнений, находить *минимакс* выражения, вычислять производные и интегралы, выводить графики сложных функций и др. Интерактивность языка обеспечивает простоту его освоения и удобство редактирования и отладки прикладных *Maple*-документов и программ. Реальная мощь языка обеспечивается не только его управляющими структурами и структурами данных, но и всем богатством его функциональных (*встроенных*, *библиотечных*, *модульных*) и прикладных (*Maple-документов*) средств, созданных к настоящему времени многими пользователями из разных прикладных областей, прежде всего математических. Важнейшим преимуществом пакета *Maple* является открытость его архитектуры, что позволило в кратчайшие сроки создать широким кругом пользователей из многих областей науки, образования, техники и т.д. обширных наборов процедур и модулей, которые значительно расширили как его возможности, так и сферу приложений. К их числу можно с полным основанием отнести и нашу библиотеку процедур, содержащую более ***640*** средств, дополняющих средства пакета, устраняющих целый ряд его *недоработок*, *расширяющих* ряд его стандартных средств и *повышающих* уровень *совместимости* релизов.

Таким образом, пакет *Maple* — не просто высоко интеллектуальный калькулятор, способный аналитически решать многие задачи, а легко обучаемая система, вклад в обучение которой вносят как сами разработчики пакета, так и его многочисленные пользователи. Очевидно, как бы ни была совершенна система, всегда найдется много специальных задач, которые

оказались за *пределами* интересов ее разработчиков. Освоив относительно простой, но весьма эффективный *Maple*-язык, пользователь может изменять уже существующие процедуры или расширять пакет *новыми*, ориентированными на решение нужных ему задач. Эти процедуры можно включать в *одну* из *пользовательских* библиотек, снабдить справочной базой, логически сцепить с главной *Maple*-пакета, так что ее средства на логическом уровне будут идентичны средствам пакета. Именно таким образом и организована наша библиотека.

Именно поэтому, в качестве основы вычислительной среды и ядра **APM** математика [141, 144, 217, 262, 286-290] мы выбрали *Maple*, представляющий на сегодня, возможно, *наиболее* мощное средство систем компьютерной систем с развитыми средствами отображения графической информации, с довольно большим набором математических функций, крупноформатных таблиц, с интерфейсом с популярным математическим пакетом ***MatLab***, с хорошо-развитым и легким в использовании графическим интерфейсом пользователя и т.д. Поддерживаемая пакетом технология, широко используется популярными средствами такими как диалоговый математический справочник (*Interactive mathematical handbook*) и диалоговый математический словарь (*Interactive mathematical dictionary*), составляющих две другие компоненты **APM**.

Создание стратегического альянса *MapleSoft Inc*. и *NAG ltd*. (*широко известного разработчика математического программного обеспечения*) имеет своей основной целью объединение усилий для создания математического программного обеспечения следующих поколений. Данный союз уже принес свои результаты, создав *новые* перспективные релизы пакета. Они позволяют пользователю определять, решать, изменять, оптимизировать, и исследовать математические задачи в любом техническом или научном проекте, включая *моделирование* и *симулирование*, теоретический анализ, техническое проектирование, и научно-прикладные разработки. Его интеллектуальные алгебраические алгоритмы, *NAG*-решатели, а также более 3500 функций различного назначения дают возможность решать весьма сложные аналитические задачи, включая продвинутый статистический анализ данных.

Наряду с вышесказанным, *Maple* имеет другие довольно развитые средства *автоматизации* и *визуализации* (*в принятой математической нотации*) *решений* многочисленных *статистических* задач. На основе *Maple* может быть создано немало интересных курсов по разделам ***"Общей Теории Статистики"***, компьютерные методические технологии и программированные курсы по отдельным разделам и темах статистического анализа данных. При этом, на основе данного пакета можно создавать достаточно удобные *Maple*-документы, решающие те или иные задачи статистической обработки данных, тогда как относительная простота пакета наряду с приемлемыми требованиями к аппаратным средствам ЭВМ выгодно отличает его от многих специализированных программных для статистической обработки данных. В то же время, пакет не ориентирован, в общем случае, на решение статистических задач и не может заменить специальные статистические пакеты. Однако, пакет может оказаться достаточно полезным средством в целом ряде статистических приложений, учитывая то обстоятельство, что он широко используется как в образовательных, так и в коммерческих целях в различных фундаментальных и прикладных областях, включая математику, физику, технику, финансы, экономику, бизнес, социологию, политические науки и т.д.

В следующем разделе мы рассмотрим основные функциональные средства *Maple*, которые ориентированы на различные статистические приложения, не останавливаясь на деталях работы в среде пакета и ограничиваясь только самыми необходимыми пояснениями. Более детально с данными вопросами читатель может ознакомиться в книгах [139-144, 190, 230, 261, 302]. В частности, книга [156] рассматривает вопросы использования пакета *Mathematica* для решения вероятностных и статистических задач; данный пакет принадлежит тому же самому классу программных средств, что и *Maple*.

## 10.5. Элементы анализа статистических данных в *Maple*

Перед рассмотрением функциональных статистических средств *Maple* вкратце поясним организационную структуру его документов (*worksheets*), представляющих своего рода программы, которые выполняются в его среде и описывают алгоритмы решаемых задач на языке *Maple*. Естественно, наше представление будет достаточно схематично; однако, оно будет вполне достаточно для понимания дальнейшего материала настоящего раздела.

*Maple*-документ это файл, описывающий, как решать математические задачи из разделов математики, науки, техники и т.д. Документы являются и диалоговыми, и многократно используемыми. Когда Вы используете *Maple*, чтобы выполнить вычисления или обработать выражения, пакет возвращает соответствующие результаты, которые Вы можете затем использовать для последующей обработки. Запрос, который Вы посылаете *Maple*, называют ***входом*** в *Maple*, а результат называют ***выходом*** *Maple*. В совокупности, ***вход*** и ***выход*** *Maple* образуютт ***выполняемую группу***, которая является базовым элементом документа. В плане структуры каждый документ представляется набором выполняемых групп.

В качестве ***выражения*** в документе выступает конструкция типа общепринятого выражения в математике, определения функции или определения процедуры, вызова функции или процедуры и т.д. Для лучшей визуализации и идентификации элементов документа *Maple*, его вход идентифицируется символом "**>**" слева, тогда как содержимое самого предложения выделяется красным цветом. Содержимое выполняемой группы отображается цветами в зависимости от информационного смысла, а именно: *выход* синим цветом, сообщения об ошибках цветом фукси и т.д. Здесь, мы дали окраску согласно стандартной установке пакета, тогда как пользователь может их переопределять по собственному усмотрению. *Выполняемая группа* идентифицируется квадратной скобкой слева, и определяет основной структурный элемент любого *Maple*-документа, который описывает алгоритм решаемой задачи вместе с результатами ее решения. Ниже будут представлены примеры различных *Maple*-документов.

Наряду с большим набором *встроенных* функциональных средств, обеспечивающих решение разнообразных математических задач, *Maple* имеет ряд специальных модулей, содержащих процедуры, ориентируемые на решение определенного класса задач. Среди этих модульных средств, есть три модуля, а именно: ***RandomTools***, ***stats*** и ***finance***, которые предназначены для решения различных задач статистического анализа данных, решения задач вероятностного характера и финансовых задач. Данные модули можно охарактеризовать следующим образом.

Модуль ***finance*** обеспечивает средства для расчетов, связанных с финансовой деятельностью. Много функций модуля вычисляют реальную стоимость различных объектов, обеспечивая упрощение ряда важных финансовых процедур. Функциональные средства данного модуля могут казаться довольно полезными при решении учетных задач и задач планирования, тесно связанных со многими прикладными статистиками.

Модуль ***RandomTools*** – набор средств для работы со случайными объектами. Эти средства могут быть также достаточно полезными при решении ряда задач статистического анализа данных. Наконец, ***stats*** модуль обеспечивает процедуры анализа данных типа различных средних и квантилей, построения различных статистических графиков, обработки данных, сглаживания данных, дисперсионного анализа и др. Данные средства могут быть полезны для решения многих задач статистического анализа данных наряду с обеспечением простого статистического анализа в целом.

Более детальный анализ данных средств показывает, что они могут оказать существенную помощь при решении в среде пакета многих статистических задач, однако в ряде случаев

они оказываются недостаточными. В таком случае пользователь должен либо полностью программировать собственный алгоритм решения статистической задачи, либо включать в него вышеупомянутые функциональные средства модулей ***stats*** и/или ***RandomTools*****,** либо использовать для своей задачи некоторый статистический пакет. В любом случае, его выбор определяется и текущей задачей, и его опытом в компьютерной технологии и статистике.

В заключении данного экскурса в *Maple* необходимо отметить, что пакет может достаточно успешно использоваться и для решения различных статистических задач, и в организации процесса обучения по курсу "*Общей Теории Статистики*"**,** что обеспечивается визуальным и интерактивным представление изучаемого материала. Встроенный язык *Maple* позволяет программировать статистические и вероятностные задачи, практически, любого уровня сложности. Наконец, на основе пакета *Maple* может быть разработан набор интерактивных практических курсов по различным разделам статистики, позволяя достаточно существенно повысить уровень практического освоения этой прикладной математической дисциплины.

Остальная часть данной главы посвящена примерам применения пакета *Maple* для решения некоторых массовых задач статистической обработки данных. Представленные процедуры обеспечивают решение некоторых важных статистических задач и могут использоваться для непосредственного практического применения в среде пакета *Maple*. Представленные ниже средства находятся в специальной пользовательской Библиотеке, описанной в нашей книге [302, 303]. Текущая версия Библиотеки содержит средства (*более* ***640*** *процедур и программных модулей*), которые ориентированы на достаточно широкую сферу приложений. Библиотека структурно подобна главной библиотеке *Maple* и снабжена детальной справочной системой. При этом, Библиотека логически связана с главной библиотекой *Maple*, обеспечивая доступ к содержащимся в ней средствам подобно стандартным средствам пакета *Maple*.

## 10.5.1. Средства для решения задач описательной статистики

Для обеспечения задач статистического анализа данных пакет *Maple* располагает целым рядом средств, поддерживаемых стандартными процедурами ***rand, randomize*** и пакетными модулями **stats** и **RandomTools**, которые содержат определения набора процедур, полезных для решения вероятностных и статистических задач. В настоящей главе представлены некоторые полезные дополнительные средства, расширяющие стандартные средства пакета, ориентированные на статистический анализ данных. Представление этих средств выполнено в разрезе четырех основных аспектов простой статистики: *описательная* статистика, элементы *регрессионного анализа*, проверка статистических гипотез и элементы *временных* (*динамических*) рядов. Данные средства представлены и процедурами, и программными модулями. В целом ряде приложений они хорошо дополняют стандартные средства пакета для решения задач подобного типа, точнее при проведении простого статистического анализа данных.

Данный раздел представляет дополнительные средства для решения задач описательной статистики. Пакет *Maple* располагает средствами для реализации статистического анализа на двух уровнях: *встроенном* и *модульном*. На *первом уровне* находится *стандартная* процедура ***rand*****,** реализующая генератор псевдослучайных чисел. Тогда как модульные средства пакета находятся в *двух* модулях ***RandomTools*** (*начиная с седьмого релиза*) и ***stats.*** Сфера приложений генератора ***rand*** весьма обширна, включая создание отладочных данных и тестовых наборов статистических данных. Однако, существенный недостаток данного средства состоит в том, что процедура ***rand*** генерирует только так называемые *нормально-подобные* распределения, которые *дополнительно* должны анализироваться. Прежде всего, это относится к получению оценки степени отклонения распределений, генерируемых процедурой ***rand*****,** от *нормального* стандартного *распределения*. Данная задача в

значительной степени решается несложной про-цедурой ***rand_Histo,*** рассматриваемой ниже и носящей вспомогательный характер.

**rand_Histo –** ***тестирование встроенного генератора псевдослучайных чисел***

Формат вызова процедуры:

**rand_Histo(a, h, n, t)**

Формальные аргументы процедуры:

**a** – целое число (*posint*)
**h** – число шагов разбиения интервала (*posint*)
**n** – величина шага разбиения
**t** – количество экспериментов (*posint*)

Описание процедуры:

С целью обеспечения общего предварительного анализа вышеотмеченной задачи создана процедура ***rand_Histo,*** которая выводит гистограмму распределения, генерируемого ***rand***-процедурой на заданном интервале. Вызов процедуры ***rand_Histo*****(a, h, n, t)** возвращает гистограмму распределения, сгенерированного процедурой ***rand*** на интервале [**a, a+h*m**]; величины **a, h** и **n** определены в описании формальных аргументов процедуры выше (**m = n** *при нечетном* **n** *и* **m = n + 1** *при четном*), а **t** определяет количество экспериментов (*количество значений*, *сгенерированных* ***rand-****процедурой*).

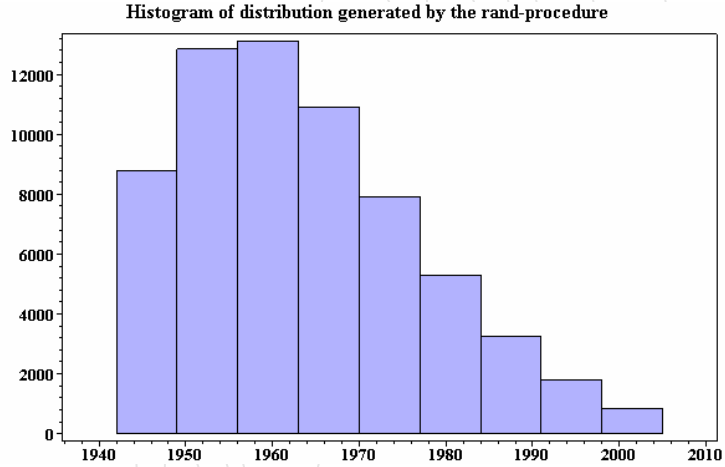
Действительно, полученная гистограмма распределения, сгенерированного процедурой ***rand*** на заданном интервале, позволяет делать вполне *определенные* заключения относительно степени и характера отклонения исследуемого распределения от *нормального* распределения. Поэтому, процедура ***rand_Histo*** вполне может рассматриваться как достаточно полезное *дополнительное* средство при использовании процедуры ***rand***. Полученная нами гистограмма распределения, сгенерированного процедурой ***rand*** на интервале [**1942 .. 2005**], позволяет делать достаточно определенные заключения относительно степени и характера отклонения исследуемого распределения от стандартного *нормального* распределения. Следует обратить внимание, что в определенной степени представленная ниже процедура ***rand_Histo*** является программным аналогом механического устройства Гальтона, иллюстрирующего *регулярность* биномиального распределения, предельным случаем для которого и является *нормальное* распределение, чрезвычайно широко используемое в статистике и теории вероятностей.

```
rand_Histo := proc(a::integer, h::integer, n::integer, t::integer)
local  omega, b, Kr, k, p, m, v;
  assign(m = `if`(type(n, 'odd'), n, n + 1)), assign(omega = rand(a .. a + h*m),
       b = [seq(0, v = 1 .. m)]);
  for k to t do for p to m do
       if  omega() <= a + p*h  and  a + (p - 1)*h < omega()  then  b[p]:= b[p] + 1;  break
       else
       end if
    end do
  end do;
  Kr:= stats['statplots', 'histogram']([seq(Weight(a + (p - 1)*h .. a + p*h, b[p]), p = 1 .. m)]);
  plots['display'](Kr, 'axes' = 'boxed', 'axesfont' = ['TIMES', 'BOLD', 10],
```

```
  'title'=`Histogram of distribution generated by the rand-procedure`, 'titlefont' = ['TIMES', 'BOLD', 12])
end proc
```

Типичный пример применения процедуры:

**> rand_Histo(1942, 7, 9, 500000);**




**Weights** – *средства работы со взвешенными данными*

**Weights_L**

**Weight_LF**

Форматы вызова процедур:

**Weights(L)**

**Weights_L(L)**

**Weight_LF(F, R, M, Z)**

Формальные аргументы процедур:

**L** – список (*list*)
**F** – символ либо строка, определяющие путь к файлу со статистическими данными
**R** – диапазон (*range*)
**M** – вычисляемое имя (*symbol*, *name*)
**Z** – вычисляемое имя (*symbol*, *name*)

Описания процедур:

Вызов процедуры ***Weights*(L)** возвращает 2-мерный массив, первая строка которого содержит отсортированное множество элементов данного статистического списка **L,** в то время как его вторая строка содержит соответствующие им веса или частоты. Тогда как вызов процедуры ***Weights_L*(L)** возвращает 2-элементную последовательность для заданного списка данных **L**. Первый элемент последовательности аналогичен результату вызова процедуры ***Weights*(L),** тогда как ее второй элемент представляет статистический список, вполне соответствующий определению статистических данных пакетного **stats**-модуля.

Наконец, вызов процедуры ***Weight_LF*(F,R,M,Z)** возвращает статистический *список* в *терминах взвешенных* данных для заданного текстового файла **F** – значений некоторого наблюдения и *интервала* **R** *допустимых* значений. Данные в **F**-файле должны состоять из значений типа {*integer*, *float*}, разделенных символами перевода каретки и возврата строки {***Enter** клавиша;* hex(0D0A)}. При этом, следует учесть, что при наличии среди значений данных *fraction*-типа

выбираются только их числители. Это обусловлено применением процедуры ***readdata*** пакета. Тогда как *замена* ее на нашу процедуру ***readdata1*** устраняет этот недостаток. Более того, через третий аргумент **M** возвращается *специальная* матрица, тогда как через четвертый **Z**-аргумент возвращается список исходных статистических данных. *Первая* строка *специальной* матрицы размерности (**2xn**) содержит *значения* элементов данных, тогда как вторая строка содержит соответствующие им *веса*.

```
Weights := proc(L::list)
local  k, p, G, R;
   assign(G = [op(sort({op(L)}))]), assign(R = array('sparse', 1 .. 2, 1 .. nops(G)));
   for k to nops(G)  do  R[1, k] := G[k];
      for p to nops(L)  do  if G[k] = L[p]  then  R[2, k] := R[2, k] + 1  end if
      end do
   end do;
   evalm(R)
end proc
Weights_L := proc(L::list)
local  k, p, G, R;
   assign(G = [op(sort({op(L)}))]), assign(R = array('sparse', 1 .. 2, 1 .. nops(G)));
   for k to nops(G)  do R[1, k] := G[k];
      for p to nops(L)  do if  G[k] = L[p]  then R[2, k] := R[2, k] + 1  end if
      end do
   end do;
   evalm(R), [seq('Weights'(R[1, k], R[2, k]), k = 1 .. nops(G))]
end proc
Weight_LF := proc(F::file, R::range({float, integer}), M::evaln, Z::evaln)
local  k, p, G, T, L, Q;
   assign(Q = readdata(F, 'float')), fclose(F), assign(Z = Q);
   L := [seq(`if`(Q[k] <= rhs(R)  and  lhs(R) <= Q[k], Q[k], NULL), k = 1 .. nops(Q))];
   assign('G' = [op(sort({op(L)}))]), assign(T = array('sparse', 1 .. 2, 1 .. nops(G)));
   for k to nops(G)  do  T[1, k] := G[k];
      for p to nops(L)  do  if G[k] = L[p]  then  T[2, k] := T[2, k] + 1  end if
      end do
   end do;
   assign(M = convert(T, 'Matrix')), [seq(Weight(T[1, k], T[2, k]), k = 1 .. nops(G))]
end proc
```

Типичные примеры применения процедур:

**> H:= [3, 8, 3, 2, 6, 3, 5, 1, 7, 1, 4, 8, 4, 6, 3, 1, 3, 2, 14, 1, 1, 1, 0, 2, 6, 6, 8, 8, 14, 14, 8, 8, 15, 7, 7, 7, 8, 8, 8, 15, 15, 15, 7, 7, 7, 8, 8, 15, 15, 8, 7, 7, 7, 7, 8, 15, 15, 15, 7]: Weights(H);**

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 14 & 15 \\ 1 & 6 & 3 & 5 & 2 & 1 & 4 & 12 & 13 & 3 & 9 \end{bmatrix}$$

**> Weights_L(H);**

$$\begin{bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 14 & 15 \\ 1 & 6 & 3 & 5 & 2 & 1 & 4 & 12 & 13 & 3 & 9 \end{bmatrix}, [Weights(0, 1), Weights(1, 6), Weights(2, 3), Weights(3, 5), Weights(4, 2), Weights(5, 1), Weights(6, 4), Weights(7, 12), Weights(8, 13), Weights(14, 3), Weights(15, 9)]$$

**> G:= rand(1 .. 63): F:= "C:\\Academy\\Examples\\RANS_IAN.dat": MkDir(F, 1): for k to 50000 do writeline(F, convert(G(), 'string')) end do: close(F); Weight_LF(F, 9 .. 16, M, Z), M;**

$$[Weigh(16., 758), Weigh(14., 821), Weigh(10., 809), Weigh(15., 781), Weigh(12., 807), Weigh(9., 769), Weigh(11., 779), Weigh(13., 825)],$$
$$\begin{bmatrix} 16. & 14. & 10. & 15. & 12. & 9. & 11. & 13. \\ 758 & 821 & 809 & 781 & 807 & 769 & 779 & 825 \end{bmatrix}$$

До некоторой степени, процедура ***Weights*** обобщает две процедуры ***statsort*** и ***frequency*** из подмодуля **transform** пакетного модуля **stats.** Наконец, процедуры ***Weights_L*** и ***Weight_LF*** могут быть полезны при работе со статистическими данными (*в терминах* **stats-***модуля*), имея ряд полезных приложений в задачах простого статистического анализа данных в среде *Maple*.

**MA** – ***вычисление скользящих средних значений***

**MAM**

**MAM1**

Форматы вызова процедур:

**MA(n, L)**

**MAM(L, n {, p})**

**MAM1(L, n {, p})**

Формальные аргументы процедур:

**n** – целое число (*integer*)
**L** – список либо текстовый файл статистических данных
**p** – (*необязательный*) положительное целое число (*posint*)

Описания процедур:

Вызов процедуры ***MA*(n, L)** возвращает список всех допустимых *скользящих средних* значений с длиной **n** скользящего сегмента для заданного списка **L** статистических данных. Тогда как вызов процедуры ***MAM*(L,n)** возвращает список всех допустимых *скользящих средних* значений с длиной **n** скользящего сегмента для заданного списка **L** статистических данных. Если задан третий *необязательный* аргумент **p,** то результат возвращается с **p** значащими цифрами.

Данные в **L**-файле должны содержать значения типа {*integer*, *float*}, разделенных символами перевода строки и возврата каретки {***Enter*** *клавиша;* hex(0D0A)}. При этом, рекомендуется обратить внимание на то обстоятельство, что при *наличии* среди значений данных *fraction*-типа процедура выбирает только их *числители*. Это обусловлено использованием в ***MAM*** процедуре *Maple*- процедуры ***readdata,*** тогда как ***MAM1*(L,n)** процедура, базирующаяся на нашей ***readata1*** процедуре, не отягощена данным (*достаточно существенным*) недостатком.

```
MA := proc(n::integer, L::{list, file})
local  a, k, R, Tb_MA;
    `if`(type(L, 'list'), assign(a = L), assign(a = readdata(L, 1))),
          assign(R = a, k = 0, Tb_MA = table());
  while  k*(n - 1) <= nops(a) - n  do  k := k + 1;
    R := stats['transform', 'moving' [n, 'mean']](R);  Tb_MA[k] := R
  end do;
```

```
    Tb_MA['k'] $ ('k' = 1 .. k)
end proc
MAM := proc(L::{list, file}, n::posint)
local  a, b, k, p;
     `if`(type(L, 'list'), assign(b = L), assign(b = readdata(L, 1)));
  assign(a = []), [seq(assign('a' = [op(a), sum(b[p], p = k .. k + n – 1)/n]), k = 1 .. nops(b) – n + 1),
     `if`(nargs = 3  and  type(args[3], 'posint')  and  2 <= args[3], evalf(op(a), args[3]), op(a))]
end proc
MAM1 := proc(L::{list, file}, n::posint)
local  a, b, k, p;
     `if`(type(L, 'list'), assign(b = L), assign(b = evalf(convert(readdata1(L, 1), 'list1'))));
     assign(a = []), [seq(assign('a' = [op(a), sum(b[p], p = k .. k + n – 1)/n]),
          k = 1 .. nops(b) – n + 1), `if`(nargs = 3  and  type(args[3], 'posint')  and  2 <= args[3],
          evalf(op(a), args[3]), op(a))]
end proc
```

Типичные примеры применения процедур:

**> A:= [3, 8, 3, 2, 6, 3, 5, 1, 7, 1, 4, 8, 4, 6, 3, 1, 3, 2, 14, 1, 1, 1, 0, 2, 6, 6, 8, 8, 14, 14, 15, 15, 14, 7, 8, 7, 8, 15]:  MA(14, A);  MAM(A, 3);**

$$\left[\frac{61}{14}, \frac{61}{14}, \frac{27}{7}, \frac{27}{7}, \frac{27}{7}, \frac{31}{7}, \frac{30}{7}, 4, 4, \frac{7}{2}, \frac{25}{7}, \frac{26}{7}, \frac{25}{7}, \frac{27}{7}, 4, \frac{67}{14}, \frac{40}{7}, \frac{46}{7}, \frac{15}{2}, \frac{15}{2}, \frac{111}{14}, \frac{59}{7}, \frac{62}{7}, \frac{66}{7}, \frac{145}{14}\right],$$

$$\left[\frac{773}{196}, \frac{192}{49}, \frac{387}{98}, \frac{200}{49}, \frac{419}{98}, \frac{127}{28}, \frac{233}{49}, \frac{983}{196}, \frac{1045}{196}, \frac{159}{28}, \frac{299}{49}, \frac{1291}{196}\right]$$

$$\left[\frac{14}{3}, \frac{13}{3}, \frac{11}{3}, \frac{11}{3}, \frac{14}{3}, 3, \frac{13}{3}, 3, 4, \frac{13}{3}, \frac{16}{3}, 6, \frac{13}{3}, \frac{10}{3}, \frac{7}{3}, 2, \frac{19}{3}, \frac{17}{3}, \frac{16}{3}, 1, \frac{2}{3}, 1, \frac{8}{3}, \frac{14}{3}, \frac{20}{3}, \frac{22}{3}, 10, 12, \frac{43}{3}, \frac{44}{3}, \frac{44}{3}, 12, \frac{29}{3}, \frac{22}{3}, \frac{23}{3}, 10\right]$$

**> AVZ:= [42, 61, 47, 56, 67, 36, 62, 40, 89, 14, 96, 7]:  MAM(AVZ, 4, 6);**

[51.5000, 57.7500, 51.5000, 55.2500, 51.2500, 56.7500, 51.2500, 59.7500, 51.5000]

**> MA(3, "C:\\Academy\\Examples\\Sample2.dat");**

[73.6666667, 74.6666667, 81.3333333, 89.3333333, 90.6666667, 87.3333333, 87.0000000, 82.3333333, 77.3333333], [76.5555555, 81.7777778, 87.1111111, 89.1111111, 88.3333333, 85.5555555, 82.2222222], [81.8148148, 86.0000000, 88.185185, 87.6666666, 85.3703704], [85.3333333, 87.2839506, 87.0740741], [86.5637860]

**> MAM("C:\\Academy\\Examples\\Sample2.dat", 3);**

[73.66666667, 74.66666667, 81.33333333, 89.33333333, 90.66666667, 87.33333333, 87.00000000, 82.33333333, 77.33333333]

**> MAM1("C:\\Temp\\Common\\Mws6789\\Data.txt", 3);**

[46.20000000, 28.80000000, 3.091666667, 1.836111111, 1.169444444, 8.544444443]

**CDiag –** ***визуализация статистической секторной диаграммы***

Формат вызова процедуры:

**CDiag(Pr, Cl, Obj)**

Формальные аргументы процедуры:

**Pr** – последовательность процентов
**Cl** – последовательность цветов
**Obj** – пояснения к выводимым объектам

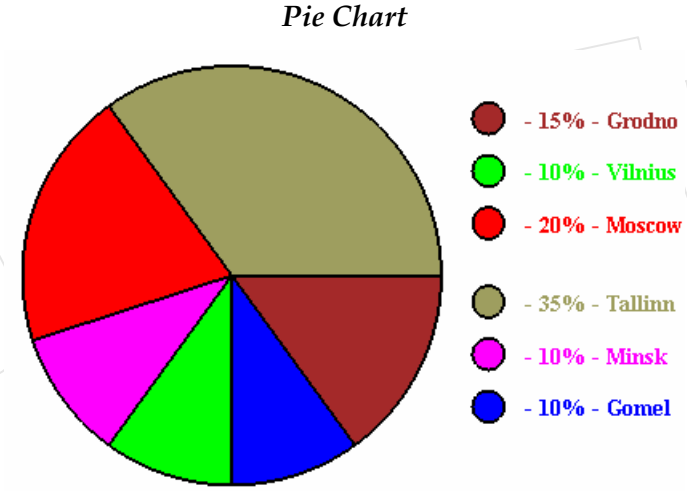
Описание процедуры:

Простая процедура ***CDiag*** на основе заданных статистических данных возвращает *секторную диаграмму*. В качестве формальных аргументов процедуры ***CDiag*(Pr, Cl, Obj)** выступают *проценты* (**Pr**), *раскраски* (**Cl**) и *пояснения* (**Obj**) представляемых диаграммой объектов. Сумма процентов не должна превысить **100** и порядок кодирования аргументов при вызове ***CDiag*** процедуры должен соответствовать указанному выше. При этом, каждое значение процента должно строго соответствовать раскраске и одному пояснению для объекта. В качестве значений раскраски допускаются имена цветов, поддерживаемые ***color***-опцией стандартной ***plot***-процедуры. При нарушении указанных требований к кодированию аргументов вызов процедуры ***CDiag*** вызывает соответствующую ошибочную ситуацию. Процедура является довольно полезным средством для быстрого вывода секторной диаграммы и для применения ее в сочетании с другими процедурами, имеющими дело со статистической обработкой.

```
CDiag := proc()
local  a, n, k, L, E;
   assign(L = [], E = proc(x)  error  "arguments are wrong %1", x  end proc, a = [args]),
      `if`(nargs = 0, E(a), assign(n = 1/3*nargs)), `if`(type(n, 'integer') <> `true`, E(a),
      `if`(100 < `+`(args[k] $ (k = 1 .. n)), E(a), [with('plots', 'display', 'textplot'),
      with('plottools', 'disk', 'pieslice')]));
   for k to n do L := [op(L), pieslice([0, 0], 4, `if`(k = 1, 0, sum(1/50*Pi*args[p],
       p = 1 .. k – 1)) .. sum(1/50*Pi*args[p], p = 1 .. k), 'color' = args[n + k]),
      disk([4.9, 0.5*(-1)^k*k], 0.3, 'color' = args[n + k]), textplot([5.5, 0.5*(-1)^k*k,
      cat(``, ` - `, args[k], `% - `, args[2*n + k])], 'align' = 'RIGHT', 'color' = args[n + k])]
   end do;
      display(L, 'scaling' = 'constrained', 'axes' = 'none', 'titlefont' = ['HELVETICA',
      'BOLDOBLIQUE', 14], 'font' = ['TIMES', 'BOLD', 11], 'title' = "Pie Chart", 'thickness' = 2)
end proc
```

Типичный пример применения процедуры:

```
> CDiag(35, 20, 10, 10, 10, 15, khaki, red, magenta, green, blue, brown, Tallinn, Moscow,
  Minsk, Vilnius, Gomel, Grodno);
```



**sHisto –** ***специальный тип оформления столбчатой диаграммы***

Формат вызова процедуры:

**sHisto(d, L::nestlist** {**, T**}**)**

Формальные аргументы процедуры:

**d** – ширина столбца
**L** – вложенный список, определяющий высоты, цвета и тексты для оснований столбцов
**T** – (*необязательный*) заголовок столбчатой диаграммы

Описание процедуры:

Пакет имеет достаточно ограниченный набор возможностей для *графического* представления статистических данных. Следовательно, в целом ряде случаев для этих целей пользователь вынужден использовать другое программное обеспечение или программировать *собственные* процедуры в *Maple*-среде. Встроенный *Maple*-язык предоставляет довольно развитую среду программирования, чтобы обеспечить для этого необходимые средства. В качестве полезного и *поучающего* примера представим ***sHisto***-процедуру, которая позволяет как оформлять, так и выводить специальный тип столбчатых статистических диаграмм.

Процедура ***sHisto*** выводит *столбчатую диаграмму* на основе *исходных* статистических данных. Обязательными аргументами при вызове процедуры ***sHisto*****(d, L)** являются: *ширина* столбца (**d**) и *вложенный* список **L,** чьи 3-элементные подсписки **[a, b, c]** определяют соответственно высоты (**a**), цвета (**b**) и тексты (**c**) для оснований столбцов создаваемой диаграммы. В качестве значений для раскраски используются имена цветов, допустимые ***color***-опцией стандартной ***plot***-процедуры. Выровненный текст (**c**) для основания столбцов располагается под столбцом, тогда как выровненное значение (**a**) его высоты располагается над столбцом. При этом, цвет (**b**) относится как к *самому* телу столбца, так и к другим элементам его оформления. Третий необязательный аргумент **T** определяет *заголовок* диаграммы; он должен иметь *string*-тип. Заголовок диаграммы использует шрифт **[*TIMES*, *BOLD*, 14]** и окрашен черным цветом.

Некорректное кодирование фактических аргументов при вызове процедуры инициирует соответствующую ошибочную ситуацию. Процедура ***sHisto*** оказывается довольно полезным средством как для быстрого вывода столбчатых *диаграмм* описанного выше оформления, так и для использования в сочетании с *другими* процедурами, имеющими дело со статистической обработкой данных. Кроме того, ее реализация использует ряд нестандартных приемов, позволяющих упрощать программирование подобных процедур в среде *Maple*. Прежде всего, это касается динамической генерации неопределенного количества графических объектов – колонок диаграммы с цветовым и текстовым оформлением. Читателю рекомендуется более детально рассмотреть данную процедуру, которая представляет также и *самостоятельный* интерес с практической точки зрения. При этом, для полного понимания текста процедуры читатель может вполне ограничиться информацией в объеме настоящей книги.

```
sHisto := proc(d, L::nestlist)
local  k, var, t1, t2, n;
    n := nops(L);  var := [seq(cat(`0`, k), k = 1 .. n)];  t1 := [seq(cat(`1`, k), k = 1 .. n)];
    t2 := [seq(cat(`2`, k), k = 1 .. n)];
    seq(assign(var[k] = plottools['rectangle']([d*(k – 1), 0], [k*d, L[k][1]], 'color' = L[k][2],
        'thickness' = 2), seq(t1[k] = plots['textplot']([1/2*(2*k – 1)*d, -0.1, convert(L[k][3], 'symbol')],
        'align' = 'BELOW', 'color' = 'black'), k = 1 .. n), seq(t2[k] = plots['textplot']([1/2*(2*k – 1)*d,
       L[k][1] + 0.1, convert(L[k][1], 'symbol')], 'align' = 'ABOVE', 'color' = 'black',
```

```
        'font' = ['TIMES', 'BOLD', 18]), k = 1 .. n)), k = 1 .. n);
    plots['display']({op(eval(t1)), op(eval(var)), op(eval(t2))}, 'tickmarks' = [0, 0],
        'font' = ['TIMES', 'BOLD', 13], 'axes' = 'none', `if`(nargs = 3  and  type(args[3], 'string'),
    op({'title' =args[3], 'titlefont' =['TIMES', 'BOLD', 18]}), NULL)), unassign(op(var), op(t1), op(t2))
end proc
```

Типичные примеры применения процедуры:

**> H:= sHisto(7, [[2, grey, 1995], [2, green, 1996], [3, yellow, 1997], [4, tan, 1998], [3, plum, 1999], [1, aquamarine, 2000], [3, coral, 2001], [6, pink, 2002], [1, turquoise, 2003], [5, gray, 2004]], "Distribution of quantities of books published during 1995 - 2004"): L:=[2, 2, 3, 4, 3, 1, 3, 6, 1, 5]:**

**> with(plots):  Lz:= listplot([seq([3.5 + (k - 1)*7, L[k]], k = 1 .. 10)], thickness = 2, color = blue):**

Warning, the name changecoords has been redefined

**> with(stats):  with(plottools):  fit[leastsquare[[x, y], y = a*x^2 + b*x + c, {a, b, c}]]([[seq(3.5 + (k - 1)*7, k = 1 .. 10)], L]);**

Warning, the names arrow and transform have been redefined

$$y = 0.0001546072981*x^2 + 0.01515151511*x + 2.217803031$$

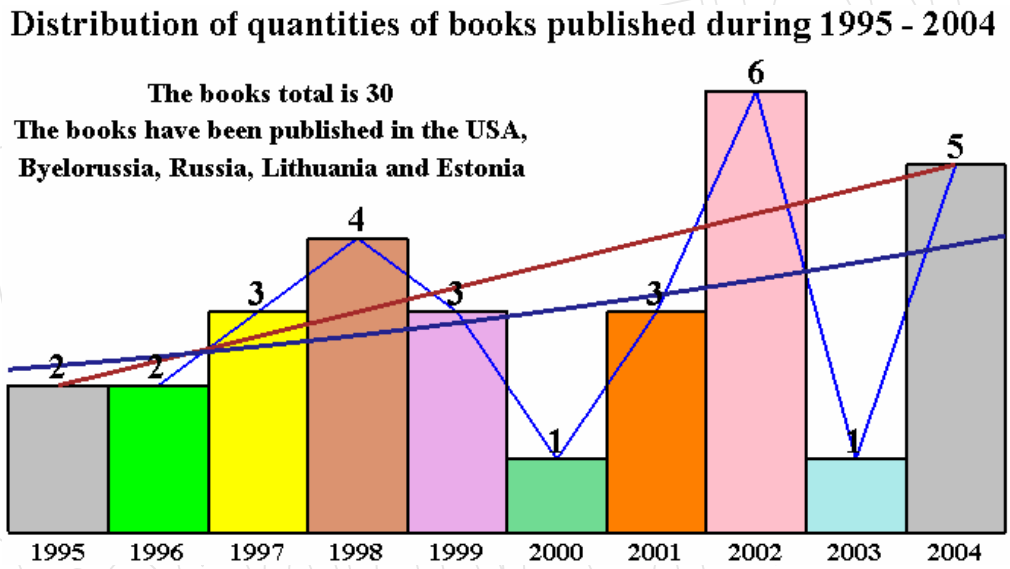
**> V:= plot(A(x, [seq(3.5 + (k - 1)*7, k = 1 .. 10)], L), x = 3.5 .. 66.5, thickness = 3, colour = orange):**

**> P:= plot(0.0001546072981*x^2+0.01515151511*x+2.217803031, x=0..70, thickness=3, color=navy):**

**> Z:= plottools[line]([3.5, 2], [66.5, 5], thickness = 3, color = brown):**

**> g:= textplot({[18.5, 6, `The books total is 30`], [18.5, 5.5, `The books have been published in the USA,`], [18.5, 5, `Byelorussia, Russia, Lithuania and Estonia`]}):**

**> plots[display]({H, Lz, P, Z, V, g}, font = [TIMES, BOLD, 14]);**




**DAAF –** ***сглаживание данных методом наименьших квадратов***

Формат вызова процедуры:

**DAAF(F::*algebraic*, x::*symbol*, da::{*Array*, *array*, *listlist*, *table*} {, 'G'})**

Формальные аргументы процедуры:

**da** – сглаживаемые данные, определенные структорой типов {*Array*, *array*, *listlist*, *table*}

**x** – символ, определяющий независимую переменную (*symbol*, *name*)

**F** – алгебраическое выражение от независимой переменной **x**
**G** – (*необязательный*) невычисленное имя (*symbol*, *name*)

Описание процедуры:

Пакет *Maple* имеет ряд средств для *сглаживания* данных. Между тем, в статистическом анализе весьма широко используется сглаживание данных посредством **МНК**. Следующая процедура обеспечивает *сглаживание* данных посредством **МНК**, основываясь на произвольной функции от одной независимой переменной.

Процедура ***DAAF*** возвращает искомую алгебраическую функцию **F(x),** которая сглаживает статистические данные, определенные структурой **da** типа {*Array*, *array*, *listlist*, *table*}. Вызов процедуры ***DAAF*(F,x,da {,G})** использует три обязательных аргумента: **F** – алгебраическое выражение, определяющее *искомую* сглаживающую функцию от одной переменной, **x** – имя не-зависимой переменной и **da** – *сглаживаемые* данные, определенные структурой **da** типа {*table*, *Array*, *array*, *listlist*}. При этом, вторая строка **da** (*Array или array*), первый подсписок **da** (*listlist*), или входы таблицы **da** определяют значения для оси абсцисс, тогда как *первая* строка **da** (*Array*, *array*), второй подсписок **da** (*listlist*), или выходы таблицы **da** определяют значения для оси ординат. Все элементы **da** должны иметь *numeric*-тип, иначе возникает ошибочная ситуация. Искомая функция **F(x)** может содержать произвольное число параметров, которые определяются в процессе сглаживания данных по методу наименьших квадратов.

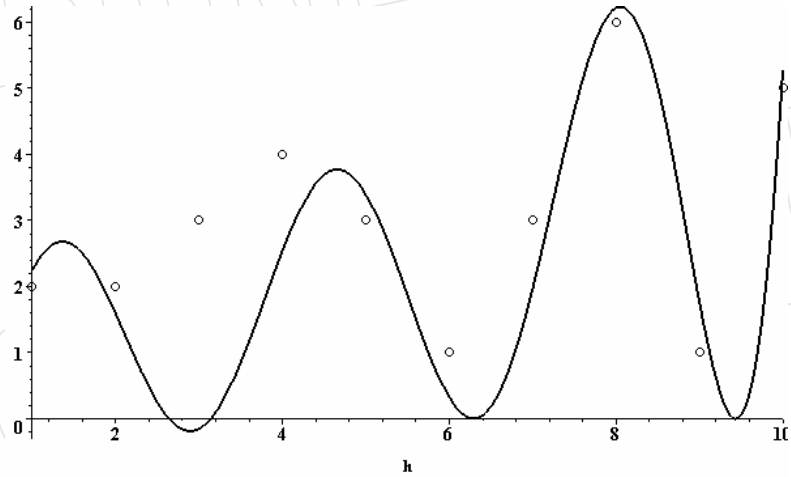
Точнее, ***DAAF*** рассматривает все *имена* выражения **F(x),** исключая переменную **x,** в качестве параметров, которые должны быть вычислены в процессе *сглаживания*. Если вызов процедуры ***DAAF*(F, x, da, G)** использует четвертый аргумент **G,** то через него возвращается совместный график точек статистических данных **da** и искомой сглаживающей кривой **F(x).** Процедура обрабатывает все основные ошибочные ситуации с выводом соответствующих сообщений. Более того, реализация процедуры использует ряд нестандартных приемов, позволяющих упрощать программирование подобных процедур. Читателю рекомендуется более подробно рассмотреть данную процедуру, которая представляет также *самостоятельный* практический интерес. При этом, для полного понимания текста процедуры читатель может ограничиться информацией в объеме настоящей книги.

```
DAAF := proc(F::algebraic, x::symbol, da::{listlist, table, Array, array})
local  a, b, c, d, k, h, p, f, n, m;
  if type(da, {'Array', 'array'})  then  assign(n = rhs(op(2, eval(da))[1]),
    m = rhs(op(2, eval(da))[2]));
    if n = 2  then  d := table([seq(da[1, k] = da[2, k], k = 1 .. m)])
    else error  "dimensionality of the third argument is invalid"
    end if
  elif type(da, 'listlist')  then  assign(n = nops(da), m = nops(da[1]));
    if n = 2  then  d := table([seq(da[1][k] = da[2][k], k = 1 .. m)])
    else error  "dimensionality of the third argument is invalid"
    end if
  else assign(d = da)
  end if;
  if op(mapTab(type, d, 2, 'numeric')) <> [`true` = `true`]
  then error  "third argument has indices and/or entries of type different from numeric"
```

```
  end if;
  assign(a = [op(indets(F(x))  minus  {x})], p = sort(map(op, [indices(d)])));
      a := [seq(`if`(type(a[k], 'symbol'), a[k], NULL), k = 1 .. nops(a))];
  if a = []  then error  "expression <%1> has no parameters", F(x)  else
      c := fsolve({seq(add((F(x) - d[x])*diff(F(x), b), x = p) = 0, b = a)}, {op(a)});
      c := {seq(`if`(type(rhs(c[k]), 'numeric'), c[k], NULL), k = 1 .. nops(c))}
  end if;
  if nops(c) <> nops(a)  then error  "set of parameters cannot be defined"
  else  h := (x) -> evalf(subs(c, F(x)))
  end if;
  if 3 < nargs  and  type(args[4], 'symbol')  then
      f := plot(h(x), x = p[1] .. p[-1], 'color' = 'blue', 'thickness' = 2,
              'labelfont' = ['TIMES', 'BOLD', 10]);
      assign(args[4] = plots['display']([f, plots['listplot'](TabList(d), 'style' = 'POINT',
              'thickness' = 5,  'symbol' = 'CIRCLE', 'symbolsize' = 14, 'color' = 'red')],
              'axesfont' = ['TIMES', 'BOLD', 11]));
      h(x)
  else
      h(x)
  end if
end proc
```

Типичные примеры применения процедуры:

```
> F := x -> (a + d*x + e*x^2 + f*x^3 + h*x^4 + g*x^5)*sin(x):  L:= array(1 .. 2, 1 .. 10, [[1, 2, 3, 4, 5, 6,
  7, 8, 9, 10], [2, 2, 3, 4, 3, 1, 3, 6, 1, 5]]):  DAAF(F, h, L, 'Z');
   (-1.241826732 + 7.420679012*h - 4.046733254*h^2 + 0.5514293876*h^3 - 0.002294331*h^5)*sin(h)
> Z;
```




Процедуры, представленные выше, могут быть полезны при ряда решении задач, имеющих дело с простой обработкой статистических данных различных характера и назначения. При этом, с учебной точки зрения они представляют не меньший интерес, используя целый ряд достаточно полезных приемов программирования в среде *Maple*-языка. В целом же, и сам

пользователь для собственных нужд решения конкретных довольно несложных прикладных задач статистического (*да и в более широком смысле*) анализа данных может запрограммировать их в среде *Maple*. В целом ряде случаев такой подход представляется даже намного более эффективным, чем использование для этих целей специальных средств анализа данных.

## 10.5.2. Средства для решения задач регрессионного анализа

Для сглаживания статистических данных и построения *моделей регрессии* двух типов (*линейной* **– Y=a*X+b** *и нелинейной* **– Y=a*X^2+b*X+c**) в статистике *весьма* широко используется **МНК**. С целью обобщения решения задачи построения *линейных* и *нелинейных* (*квадратичных*) *моделей* регрессии для заданных результирующей и факторной переменных, а также для вычисления *коэффициентов корреляции* (**CC**) и *отношения корреляции* (**CR**) с графическим представлением в *единой* системе координат *исходных* данных и модели *регрессии* в среде пакета *Maple* может также использоваться специальная ***LRM_NRM***-процедура. Процедура оказывается достаточно полезным средством для проведения быстрого анализа оценки степени корреляции данных в среде *Maple*.

**LRM_NRM –** ***однофакторные модели регрессии***

Формат вызова процедуры:

**LRM_NRM(A, B, T, P, CR)**

Формальные аргументы процедуры:

**A** – список, вектор либо текстовый файл данных (*результирующая переменная*)
**B** – список, вектор либо текстовый файл данных (*факторная переменная*)
**T** – тип искомой модели регрессии (*может быть* ***LRM*** *либо* ***NRM***)
**CR** – вычисляемое имя (*symbol*, *name*)
**P** – вычисляемое имя (*symbol*, *name*)

Описание процедуры:

Для сглаживания статистических данных и построения вышеупомянутых моделей *регрессии* служит специальная процедура ***LRM_NRM*(A, B, T,P,CR)** с пятью *формальными* аргументами. Формальные аргументы ***LRM_NRM***-процедуры имеют следующее назначение. Первый и второй фактические аргументы **A** и **B** определяют *списки*, *векторы* или текстовые *файлы* для результатной и факторной переменных соответственно. Третий аргумент **T** определяет тип разыскиваемой модели регрессии (***LR***M – y = a*x+b, ***NRM*** – y = a*x^2+b*x+c). Наконец, через четвертый и пятый фактические аргументы **P** и **CR** возвращаются соответственно *единый* график распределения исходных статистических данных и кривой регрессии (*т. е. модели регрессии*), и корреляционное отношение.

Данные в файлах **A** и **B** должны быть значениями типа {*integer*, *float*, *fraction*}, разделенных символами перевода строки и возврата каретки {***Enter*** *клавиша*; hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с разными длинами производится их выравнивание по минимальной длине с выводом соответствующего информационного сообщения.

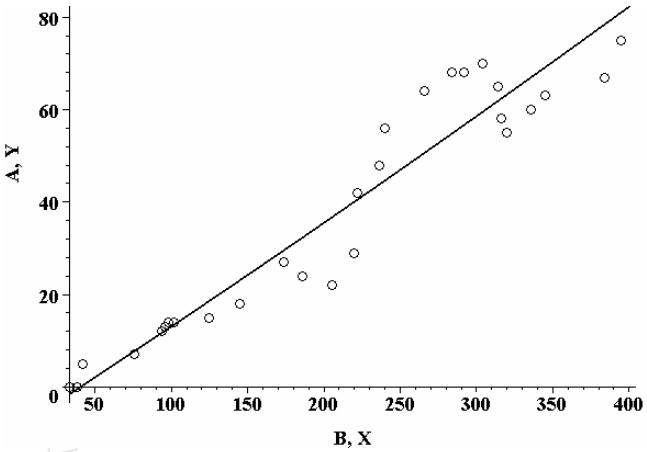
Процедура ***LRM_NRM*** непосредственно возвращает искомую *однофакторную модель регрессии*. Следующий фрагмент представляет пример применения процедуры для решения задачи создания линейной и нелинейной моделей регрессии для *результатной* **U**-переменной и *факторной* **A**-переменной, значения для которых были выбраны согласно числу ежегодных отечественных и зарубежных цитирований наших научных работ по математической теории однородных структур (*Cellular Automata*) и кибернетике в целом.

```
LRM_NRM := proc(A::{vector, file, list}, B::{vector, file ,list}, T::name, P::evaln, CR::evaln)
local  a, b, k, m, n, p, L, M, N, r, omega, o, G, X, Y, P1, P2, F, sr, ds, v;
  `if`(type(A, 'list'), assign(a = A), `if`(type(A, 'vector'), assign(a = convert(A, 'list1')),
      assign(a = map(op, convert(readdata1(A, 1), 'list1')))));
  `if`(type(B, 'list'), assign(b = B), `if`(type(B, 'vector'), assign(b = convert(B, 'list1')),
      assign(b = map(op, convert(readdata1(B, 1), 'list1')))));
  assign(o = {nops(b), nops(a)}), `if`(o[1] < 3, ERROR("One of samples is too small"),
      [assign('a' = a[1 .. o[1]], 'b' = b[1 .. o[1]]), `if`(1 < nops(o), WARNING("Lengths of
      samples is different, a levelling by minimal length has been done"), NULL)]);
  `if`(nops(a) <> nops(b), ERROR("different dimensionalities of lists of source statistical
      data"), assign(omega = [seq(1, p = 1 .. nops(a))], G = [], F = ['TIMES', 'BOLD', 10]));
  assign(v = (() -> op([assign('L' = []), seq(assign('L' = [op(L), product(args[m][p],
      m = 1 .. nargs)]), p = 1 .. nops(args[1])), L])));
  assign(sr = (() -> `+`(args)/nargs), ds = (() -> `+`(seq((args[k] - sr(args))^2,
      k = 1 .. nargs))/nargs)), `if`(T = 'LRM', assign('M' = Matrix(2, 2, [[sr(op(v(b, b))),
      sr(op(b))], [sr(op(b)), 1]]), 'N' = Vector(2, [sr(op(v(a, b))), sr(op(a))])), `if`(T = 'NRM',
      assign('M' = Matrix(3, 3, [[sr(op(v(b, b, b, b))), sr(op(v(b, b, b))), sr(op(v(b, b)))],
      [sr(op(v(b, b, b))), sr(op(v(b, b))), sr(op(b))], [sr(op(v(b, b))), sr(op(b)), 1]]),
      'N' = Vector(3, [sr(op(v(a, b, b))), sr(op(v(a, b))), sr(op(a))])),
      ERROR("regression model type <%1> is inadmissible", T)));
  assign('r' = convert(evalf(LinearAlgebra:- LinearSolve(M, N)), 'list1'));
  assign(Y = `if`(nops(r) <> 1,
     proc(X)  local n; sum(r[n]*X^(nops(r) – n), n = 1 .. nops(r))  end proc,
     proc(X)  sign(op(r))*X  end proc)), `if`(nops(r) = 1, assign('CR' = sign(op(r))), NULL);
  `if`(nops(r) = 2, assign('CR' = evalf(sqrt((ds(op(a)) – sr(op(v(a – r[1]*b – r[2]*omega,
      a – r[1]*b – r[2]*omega)))))/ds(op(a))))), `if`(nops(r) = 3, assign('CR' = evalf(sqrt((ds(op(a)) –
      sr(op(v(a – r[1]*v(b, b) – r[2]*b – r[3]*omega, a – r[1]*v(b, b) –
      r[2]*b – r[3]*omega)))))/ds(op(a))))), NULL));
  seq(assign('G' = [op(G), [b[n], a[n]]]), n = 1 .. nops(a));
  P1 := plots['pointplot'](G, 'color' = 'blue', 'thickness' = 3, 'symbol' = 'CIRCLE', 'symbolsize' = 16);
  P2 := plot(Y(X), X = b[1] .. b[-1], 'thickness' = 2, 'color' = 'green');
  assign(P = plots['display']([P1, P2], 'axesfont' = F, 'scaling' = 'UNCONSTRAINED', 'labels' =
      ["B, X", "A, Y"], 'labeldirections' = ['HORIZONTAL', 'VERTICAL'], 'labelfont' = F)), sort(Y(X))
end proc
```

Типичные примеры применения процедуры:

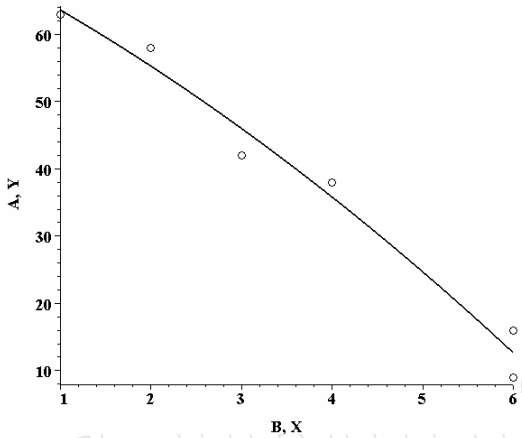
**> U:= [0, 0, 5, 7, 12, 13, 14, 14, 15, 18, 27, 24, 22, 29, 42, 48, 56, 64, 68, 68, 70, 65, 58, 55, 60,63,67,75, 95]:**
**A:= [33, 38, 42, 76, 94, 96, 98, 102, 125, 145, 174, 186, 205, 220, 222, 236, 240, 266, 284, 292, 304, 314, 316, 320, 336, 345, 384, 395, 405]: LRM_NRM(U, A, NRM, *Figure*, CR), CR; *Figure;***

0.00003166933948*X^2 + 0.2148197209*X - 8.714233847,  0.9582099422




**> LRM_NRM("C:\\Academy\\Examples/Members.dat", "C:\\Academy\\Examples/Age.dat", NRM, Figure, CR), CR; Figure;**

**-**0.4611360239*X^2 **-** 6.975336323*X **+** 71.08221226**,** 0.9886930852




В результате выполнения ***LRM_NRM***-процедуры с заданными значениями для ее аргументов было получено значение **CR(U, A)=0.9582099422** для отношения корреляции. Через аргумент ***Figure*** процедуры возвращен *общий* график отыскиваемой *квадратичной* модели *регрессии* и *распределения* точек *исходных* статистических данных (**U,A**). Полученные значения отношения *корреляции* как на основе ***LRM,*** так и на основе ***NRM*** говорят о наличии достаточно тесной связи между переменными **U** и **A** статистического наблюдения.

**lsf –** ***сглаживание статистических данных методом наименьших квадратов***

Формат вызова процедуры:

**lsf(A, x, F, S::*evaln*)**

Формальные аргументы процедуры:

**A** – (**2xn**) *NAG*-массив с числовыми элементами (*сглаживаемые данные*)
**x** – символ (*имя ведущей переменной алгебраического выражения*)
**F** – алгебраическое выражение от ведущей **x** переменной
**S** – вычисляемое имя (*symbol*, *name*)

Описание процедуры:

Для сглаживания данных методом наименьших квадратов может быть полезна и процедура ***lsf*(A, y, x, F, S).** Она возвращает уравнение, определяющее искомую кривую, вычисленную на основе заданного алгебраического выражения **F** и точек *исходных* данных, заданных *NAG*-

массивом размерности (**2xn**) с числовыми элементами. *Первая* и *вторая* строки *NAG*-мас-сива **A** определяют значения факторной и результатной переменной соответственно. Тогда как через переменную **S** процедура возвращает общий график найденной *сглаживающей* кривой и распределение точек исходных данных **A.** В случае невозможности вычислить параметры заданного выражения **F** процедура возвращает *NULL*-значение, с выводом соответствующего сообщения. Возвращаемое сглаживающее выражение не поддерживает *прямого* вычисления его значений в требуемых точках. Для этого может использоваться, например, следующая конструкция ***subs*(x = b, <*выражение*>),** где **b** – требуемая точка. Процедура ***lsf*** имеет целый ряд достаточно полезных приложений в статистическом анализе данных.

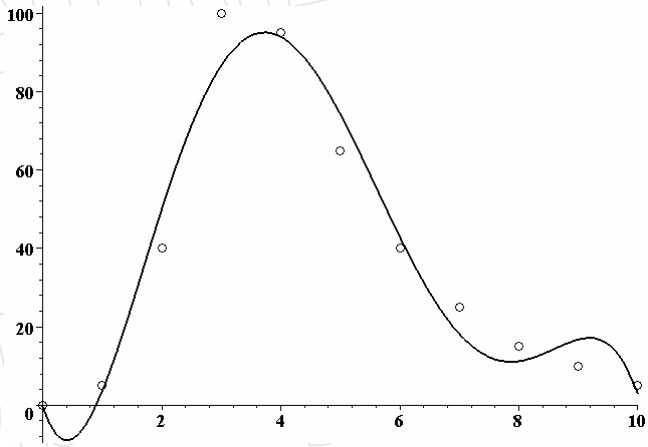
```
lsf := proc(A::Array(numeric), x::symbol, F::algebraic, S::evaln)
local  a, b, p, k, j, v, h, g, y;
  assign(a = rhs([ArrayDims(A)][2]), h = indets(F)  minus  {x});
  h := {seq(`if`(type(h[k], 'symbol'), h[k], NULL), k = 1 .. nops(h))};
  b := plots['pointplot']([seq([A[1, v], A[2, v]], v = 1 .. a)], 'symbol' = 'circle', 'symbolsize' = 14);
  g := fsolve({seq(diff(evalf(normal(simplify(expand(add((A[2, j] – subs(x = A[1, j], F))^2,
       j = 1 .. a))))), h[p]) = 0, p = 1 .. nops(h))}, h, 'fulldigits');
  try  assign(y = ((x) -> eval(subs(g, F)))), eval(subs(g, F)), assign(S = plots['display'](b, plot(y(x),
       x = A[1, 1] .. A[1, a]), 'thickness' = 2, 'axesfont' = ['TIMES', 'BOLD', 9]))
  catch  "wrong number (or type) of parameters":
       WARNING("parameters %1 cannot be evaluated", h)
  end try
end proc
```

Типичные примеры применения процедуры:

```
> A:= Array(1 .. 2, 1 .. 11, [[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10], [0, 5, 40, 100, 95, 65, 40, 25, 15, 10, 5]]):
  lsf(A, x, a*x^5 + b*x^4 + c*x^3 + d*x^2 + h*x + p, H);  subs(x = 63, %);  H;
```

-0.0817240592*x^5 + 2.16401955*x^4 - 19.6466462*x^3 + 66.3093028*x^2 - 44.8785384*x - 0.36810604

-51668458.24




Процедуры данного раздела могут быть достаточно полезны при решении прикладных задач регрессионного анализа и сглаживания статистических данных. Более того, с учебной точки зрения они представляют не меньший интерес, используя целый ряд довольно полезных приемов программирования в среде *Maple*-языка пакета такого типа задач.

## 10.5.3. Средства для проверки статистических гипотез

Проверка *достоверности* гипотез играет чрезвычайно важную роль в задачах статистического анализа. В настоящем разделе представлен ряд полезных процедур, решающих проблему проверки так называемой *нулевой гипотезы* (**Ho**-*гипотезы*) и на основе *параметрических* тестов (*Фишера и Стьюдента*), и на основе *непараметрических* (*Манна-Уитни и Ван дер Ваардена*).

**F_test_Ds –** ***F-тест Фишера проверки Ho-гипотезы***

Формат вызова процедуры:

**F_test_Ds(L1, L2, n)**

Формальные аргументы процедуры:

**L1** – список либо текстовый файл статистических данных (*малая выборка*)
**L2** – список либо текстовый файл статистических данных (*малая выборка*)
**n** – положительное целое число (*точность вычислений*)

Описание процедуры:

С целью упрощения проверки *достоверности* **Ho**-гипотезы относительно *равенства дисперсий* выборок на основе **F**-теста Фишера предлагается простая процедура ***F_test_Ds*(L1, L2, n),** которая возвращает список вида **[k1, k2, Ff].** В качестве формальных аргументов процедуры выступают списки или текстовые файлы **L1** и **L2** значений элементов исследуемых выборок и заданная точность вычислений **n**. Элементы возвращаемого *списка* имеют следующий смысл: (**1**) количество степеней свободы **k1** для большей дисперсии, (**2**) количество степеней свободы **k2** для меньшей дисперсии и (**3**) **Ff**-значение, вычисленное с **n**-точностью.

Данные в файлах **L1** и **L2** должны быть значениями типов {*integer,float,fraction*}, разделенными символами перевода строки и возврата каретки {***Enter** клавиша*; hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с разными длинами производится их выравнивание по длине.

Для иллюстрации применения ***F_test_Ds*(L1, L2, n)** процедуры, созданной на основе **F**-теста Фишера, посредством встроенного генератора ***rand*** псевдослучайных чисел создаются две малые выборки **L1** и **L2** объемом **n=20** и **n=22** соответственно. Вызов процедуры ***F_test_Ds*(L1, L2, n),** примененный к данным выборкам, возвращает список **[k1 = 22, k2 = 20, Ff = 1.082],** определяя количества степеней свободы **k1=22, k2=20** и значение **Ff=1.082.** Затем, на основе таблицы (*см., статистические таблицы, например* [223,225]) *критических* точек **F**-критерия для доверительного уровня **a = 5 %** (**a = 1 %**) и для количеств степеней свободы **k1=22** и **k2=20,** мы легко получаем значения **Fst = 2.07** (**Fst = 2.83**) *критических* точек соответственно. Так как для обоих доверительных уровней имеет место соотношение **Ff <<Fst,** то **Ho**-гипотеза может быть принята на высоком *доверительном* уровне и различие между *дисперсиями* обоих исследуемых выборок вполне можно полагать случайным.

Процедура ***F_test_Ds*** может быть достаточно легко *модернизирована* посредством погружения в ее тело соответствующей ссылки на таблицу (*предварительно сохраненную в файле, например, **m**-формата*) критических точек **F**-критерия. Это позволит получать ответы в терминах *true* (**Ho**-*гипотеза принимается*) и *false*-значения (**Ho**-*гипотеза отвергается*).

```
F_test_Ds := proc(L1::{list, file}, L2::{list, file}, n::integer)
local  a, b, d, c, p, k1, k2, R, L, Sr, Ds, Ff, x, y, o;
   `if`(type(L1, 'list'), assign(x = L1), assign(x = map(op, convert(readdata1(L1, 1), 'list1'))));
```

```
  `if`(type(L2, 'list'), assign(y = L2), assign(y = map(op, convert(readdata1(L2, 1), 'list1'))));
  assign(o = {nops(y), nops(x)});
  if  o[1] < 3  then  error  "One of samples is too small"  end if;
  assign67(Sr = (() -> `+`(args)/nargs), a = op(x), b = op(y), d = nops(x), c = nops(y));
  Ds := () -> sum((args[p] – Sr(args))^2, p = 1 .. nargs)/(nargs – 1);
  R := `if`(Ds(b) <= Ds(a), [k1 = d, k2 = c, Ff = Ds(a)/Ds(b)],
      [k1 = c, k2 = d, Ff = Ds(b)/Ds(a)]);  [R[1], R[2], evalf(R[3], n)]
end proc
```

Типичные примеры применения процедуры:

**> L1, L2:= [], []: Kr:= rand(42 .. 99): Art:= rand(47 .. 99): seq(assign('L1'= [op(L1), Kr()]), j = 1 .. 20);**
**seq(assign('L2'= [op(L2), Art()]), j=1 .. 22);  F_test_Ds(L1, L2, 4);** ⇒ [k1 = 22, k2 = 20, Ff = **1.082**]

**> F_test_Ds(L1, "C:\\Academy\\Examples\\Sample2.dat", 4);** ⇒ [k1 = 20, k2 = 22, Ff = **1.066**]

**T_test_AV** – ***t-критерий Стьюдента проверки Но-гипотезы***

Формат вызова процедуры:

**T_test_AV(L1, L2, n)**

Формальные аргументы процедуры:

**L1** – список либо текстовый файл статистических данных (*малая выборка*)
**L2** – список либо текстовый файл статистических данных (*малая выборка*)
**n** – положительное целое число (*точность вычислений*)

Описание процедуры:

С целью упрощения оценки достоверности **Но**-гипотезы относительно равенства *средних* совокупностей предлагается достаточно простая процедура ***T_test_AV*(L1, L2, n)**, которая возвращает список вида **[k, tf].** В качестве формальных аргументов процедуры выступают списки или текстовые файлы **L1** и **L2** значений элементов исследуемых *выборок* и заданная **n**-точность вычислений. Элементы возвращаемого списка имеют следующий смысл: (**1**) число степеней свободы **k** и (**2**) **tf**-значение, вычисленное с **n**-точностью.

Данные в файлах **L1** и **L2** должны быть значениями типов {*integer*, *float*, *fraction*}, разделенных символами перевода строки и возврата каретки {***Enter** клавиша***;*** hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с разными длинами производится их выравнивание по длине.

Для иллюстрации применения ***T_test_AV*(L1, L2, n)** процедуры, созданной на основе **T**-теста Стьюдента, посредством встроенного генератора ***rand*** псевдослучайных чисел создаются две малые выборки **L1** и **L2** объемом **n = 20** и **n = 22** соответственно. Вызов ***T_test_AV*(L1, L2, n)** процедуры, примененный к данным *выборкам*, возвращает список **[k=40, tf=0.7689],** определяя число *степеней свободы* **k=40** и значение **tf=0.7689.** Затем, на основе таблицы (*статистические таблицы*, *например* [223, 225]) критических точек **T**-критерия для доверительного уровня **a = 0.1%** (**a = 1 %**) и для числа степеней свободы **k = 40** мы легко получаем значение **tst = 3.55** критической (*стандартной*) точки. Так как для обоих *доверительных* уровней имеем соотношение **tf << tst,** то **Но**-гипотеза может быть принята на весьма *высоком* доверительном уровне (**P = 0.999**) и различие между *средними* обоих исследуемых выборок можно считать случайным.

Процедура ***T_test_AV*** может быть достаточно легко *модернизирована* посредством погружения в ее тело соответствующей ссылки на таблицу (*предварительно сохраненную в файле*, *например*,

***m**-формата*) критических точек **F**-критерия. Это позволит получать ответы в терминах *true* (**Ho**-*гипотеза принимается*) и *false*-значения (**Ho**-*гипотеза отвергается*).

```
T_test_AV := proc(L1::{list, file}, L2::{list, file}, n::integer)
local  a, b, p, h, k, L, H, Sr, Ds, Sd, tf, x, y, o;
    `if`(type(L1, 'list'), assign(x = L1), assign(x = map(op, convert(readdata1(L1, 1), 'list1'))));
    `if`(type(L2, 'list'), assign(y = L2), assign(y = map(op, convert(readdata1(L2, 1), 'list1'))));
    assign(o = {nops(y), nops(x)});
  if  o[1] < 3  then error  "One of samples is too small"  end if;
    assign(Sr = ((a)  ->  `+`(op(a))/nops(a)), a = nops(x), b = nops(y));
  Ds := (c, d)  ->  sum((c[p] – Sr(c))^2, p =1 .. nops(c)) + sum((d[h] – Sr(d))^2, h =1 .. nops(d));
  if a = b  then  Sd := sqrt(Ds(x, y)*a/(a – 1))
  else  Sd := sqrt(Ds(x, y)*(a + b)/((a + b – 2)*a*b))
  end if;
    [k = a + b – 2, tf = abs(evalf((Sr(x) – Sr(y))/Sd, n))]
end proc
```

Типичные примеры применения процедуры:

**> L1, L2:= [], []:  Kr:= rand(42 .. 99): Art:= rand(47 .. 99): seq(assign('L1'= [op(L1), Kr()]), j = 1 .. 20): seq(assign('L2' = [op(L2), Art()]), j = 1 .. 22):  T_test_AV(L1, L2, 4);**  ⇒ [k = 40, tf = **0.7689**]

**> T_test_AV(L1, "C:\\Academy\\Examples\\Sample2.dat", 4);**  ⇒ [k = 40, tf = **1.784**]

Представленные выше процедуры базируются на так называемых параметрических тестах Фишера и Стьюдента проверки **Ho**-гипотезы, тогда как следующие процедуры базируются на непараметрических критериях Ван дер Ваардена и Манна-Уитни.

**U_test_MW – *U-критерий Манна-Уитни проверки Ho-гипотезы***

Формат вызова процедуры:

**U_test_MW(L1, L2, n)**

Формальные аргументы процедуры:

**L1** – список либо текстовый файл статистических данных (*малая выборка*)
**L2** – список либо текстовый файл статистических данных (*малая выборка*)
**n** – положительное целое число (*точность вычислений*)

Описание процедуры:

С целью упрощения *проверки* достоверности **Ho**-гипотезы относительно двух *малых* выборок на основе непараметрического **U**-теста Манна-Уитни предлагается ***U_test_MW*(L1, L2, n)** процедура, возвращающая список вида **[n1, n2, Uf].** В качестве формальных аргументов процедуры выступают списки или текстовые файлы **L1** и **L2** *значений* элементов исследуемых выборок и заданная *точность* вычислений **n**. Элементы возвращаемого списка имеют смысл: (**1**) *количество* элементов (**n1**) *первой* выборки **L1,** (**2**) *количество* элементов (**n2**) *второй* выборки **L2** и (**3**) **Uf**-значение **U**-критерия, вычисленное с **n**-точностью.

Данные в файлах **L1** и **L2** должны быть значениями типов {*integer*, *float*, *fraction*}, разделенных символами перевода строки и возврата каретки {***Enter** клавиша*; hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с разными длинами производится их выравнивание по длине.

Для иллюстрации применения ***U_test_MW*(L1,L2,n)** процедуры, созданной на основе **U**-теста Манна-Уитни, посредством встроенного генератора ***rand*** псевдослучайных чисел создаются две малые выборки **L1** и **L2** объемом **n1=9** и **n2=11** соответственно. Вызов ***U_test_MW*(L1, L2, 5)** процедуры**,** примененный к данным выборкам, возвращает список вида **[n1 = 9, n2 = 11, Uf = 29.],** определяя число элементов обеих выборок и фактическое **Uf**-значение для **U**-критерия. Затем, на основе таблицы (*статистические таблицы*, *например* [223,225]) критических точек **U**-критерия мы легко получаем значение **Ust=19** для чисел **n1=9, n2=11** доверительного уровня в **1%.** Так как имеет место соотношение **Uf=29 >> Ust=19,** то **Ho**-гипотеза может быть принята с указанным доверительным уровнем **1%,** т.е. различие между *обоими* исследуемыми малыми выборками вполне можно полагать случайным.

Процедура ***U_test_MW*** может быть достаточно легко модернизирована путем погружения в ее тело соответствующей ссылки на таблицу (*предварительно сохраненную в файле*, *например*, ***m****-формата*) *критических* точек **U**-критерия. Это позволит получать ответы в *true*-значения (**Ho**-*гипотеза принимается*) и *false*-значения (**Ho**-*гипотеза отвергается*).

```
U_test_MW := proc(L1::{list, file}, L2::{list, file}, n::integer)
local  S, n1, n2, G, Srt, k, Z, r1, r2, z, x, y, N, L, a, b, o;
    `if`(type(L1, 'list'), assign(a = L1), assign(a = map(op, convert(readdata1(L1, 1), 'list1'))));
    `if`(type(L2, 'list'), assign(b = L2), assign(b = map(op, convert(readdata1(L2, 1), 'list1'))));
    assign(o = {nops(b), nops(a)});
  if  o[1] < 3  then error  "One of samples is too small"  end if;
    assign(G = proc(z)  options operator, arrow; `if`(whattype(z) = `*`, op(1, z), z)  end proc,
    Srt = proc(x, y)  options operator, arrow; `if`(G(x) <= G(y),`true`, `false`)  end proc);
    assign(L = [seq(`if`(a[k] = 1, 1.0, a[k])*S, k = 1 .. nops(a))]);
    assign(r1 = 0, r2 = 0, Z = sort([op(L), op(b)], Srt));
  for k to nops(Z)  do
    if whattype(Z[k]) = `*`  then  r1 := r1 + k  else  r2 := r2 + k  end if
  end do;
    assign(n1 = 1/2*nops(a)*(nops(a) + 1), n2 = 1/2*nops(b)*(nops(b) + 1));
    [nops(a), nops(b), evalf(min(r1 – n1, r2 – n2), n)]
end proc
```

Типичные примеры применения процедуры:

```
> L1:= [64, 68, 70, 72, 75, 76, 79, 80, 83]:  L2:= [60, 60, 62, 66, 68, 69, 70, 71, 73, 78, 80]:
  U_test_MW(L1, L2, 5);    ⇒   [9, 11, 29.]
> U_test_MW(L1, "C:\\Academy\\Examples\\Sample2.dat", 4);   ⇒   [9, 11, 31.]
```

**X_test_VW –** ***X-критерий Ван дер Ваарден проверки Но-гипотезы***

Формат вызова процедуры:

**X_test_VW(L1, L2, n)**

Формальные аргументы процедуры:

**L1** – список либо текстовый файл статистических данных (*малая выборка*)
**L2** – список либо текстовый файл статистических данных (*малая выборка*)
**n** – положительное целое число (*точность вычислений*)

Описание процедуры:

С целью упрощения проверки достоверности **Ho**-гипотезы относительно двух *малых* выборок на основе *непараметрического* **X**-теста Ван дер Ваардена предлагается ***X_test_VW*(L1, L2, n)**-процедура**,** возвращающая 2-элементную последовательность следующего вида **N, [R/(N+1)].** Аргументами процедуры являются списки или текстовые файлы **L1** и **L2** значений элементов исследуемых выборок и заданная точность вычислений **n**. Возвращаемая последовательность определяет: (**1**) количество элементов *сравниваемых* выборок (**N**), (**2**) список значений **R/(N+1),** вычисленных с **n**-точностью.

Данные в файлах **L1** и **L2** должны быть значениями типов {*integer*, *float*, *fraction*}, разделенных символами перевода строки и возврата каретки {***Enter*** *клавиша***;** hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с разными длинами производится их выравнивание по длине.

Для иллюстрации применения процедуры ***X_test_VW*(L1, L2, n),** созданной на *основе* **X**-теста Ван дер Ваардена, используются две малые выборки **L1** и **L2.** Вызов ***X_test_VW*(L1, L2, 3),** примененный относительно этих выборок, возвращает последовательность, чей первый элемент определяет *общее количество* элементов *обоих* выборок и *список* значений **R/(N+1)**

**N = 20, [0.190, 0.286, 0.429, 0.571, 0.667, 0.714, 0.810, 0.857, 0.952]**

Затем, на основе таблицы значений **ψ**-функции для *каждого* элемента списка, возвращаемого процедурой ***X_test_VW*** в качестве второго элемента, мы вычисляем значения **ψ[R/(N+1)]:**

**ψ(0.190) = -0.88, ψ(0.286) = -0.57, ψ(0.429) = -0.18, ψ(0.571) = +0.18, ψ(0.667) = +0.43,**

**ψ(0.714) = +0.57, ψ(0.810) = +0.88, ψ(0.857) = +1.07, ψ(0.952) = +1.66**

Суммируя полученные результаты, мы получаем значение $\mathbf{Xf = \Sigma_{\psi}[R/(N+1)] = 3.16.}$ Затем, для числа **N = n1 + n2 = 9 + 11=20** и *доверительного* уровня **5%** с учетом разности **n1 - n2 = 11 - 9 =2** в специальной таблице (*например*, *в* [223, 225]) критических точек **Xst** для **X**-критерия легко получаем значение **Xst=3.84.** В виду же соотношения **Xf = 3.16 < Xst = 3.84** мы имеем вполне достаточное основание принять **Ho**-гипотезу с указанным доверительным уровнем в **5%.**

Процедура ***X_test_VW*** может быть достаточно легко модернизирована путем погружения в ее тело соответствующей ссылки на таблицу (*предварительно сохраненную в файле*, *например*, ***m****-формата*) *критических* точек **Xst X**-критерия. Это позволит получать ответы в терминах *true*-значения (**Ho**-*гипотеза принимается*) и *false*-значения (**Ho**-*гипотеза отвергается*).

```
X_test_VW := proc(L1::{list, file}, L2::{list, file}, n::integer)
local  S, L, Lx, Gal, Sr, k, h, omega, z, x, y, N, a, b, o;
     `if`(type(L1, 'list'), assign(a = L1), assign(a = map(op, convert(readdata1(L1, 1), 'list1'))));
     `if`(type(L2, 'list'), assign(b = L2), assign(b = map(op, convert(readdata1(L2, 1), 'list1'))));
     assign(o = {nops(a), nops(b)});
  if o[1] < 3  then error  "One of samples is too small"  end if;
  Gal := (z)  ->  `if`(whattype(z) = `*`, op(1, z), z);  Sr := (x, y) -> `if`(Gal(x) <= Gal(y), `true`, `false`);
  if  nops(a) <= nops(b)  then  L := a; Lx := b else L := b; Lx := a  end if;
     assign('L' = [seq(`if`(a[k] = 1, 1.0, a[k])*S, k = 1 .. nops(a))]);
     assign(omega = [], h = sort([op(L), op(Lx)], Sr));
  for k to nops(h)  do
     if  whattype(h[k]) = `*`  then  omega := [op(omega), k/(nops(h) + 1)]  end if
```

```
  end do;
    N = nops(h), evalf(omega, n)
end proc
```

Типичные примеры применения процедуры:

```
> L:= [64, 68, 70, 72, 75, 76, 79, 80, 83]:  P:= [60, 60, 62, 66, 68, 69, 70, 71, 73, 78, 80]:
  X_test_VW(L, P, 3);
                    N = 20, [0.190, 0.286, 0.429, 0.571, 0.667, 0.714, 0.810, 0.857, 0.952]
> X_test_VW(L, "C:\\Academy\\Examples\\Sample2.dat", 4);
                  N = 20, [0.1429, 0.1905, 0.2381, 0.2857, 0.4286, 0.4762, 0.5238, 0.6190, 0.7143]
```

Представленные в разделе процедуры имеют ряд полезных приложений в статистическом анализе данных. Как уже отмечалось, процедуры могут быть модернизированы посредством имплантации в их тело соответствующей ссылки на таблицу (*предварительно сохраненную в* ***m****-файле*) критических точек. Это позволит получать ответы в терминах *true*-значения (**Ho**-*гипотеза принимается*) и *false*-значения (**Ho**-*гипотеза отвергается*). Оставляем это читателю в качестве весьма полезного практического упражнения по программированию в среде *Maple*-языка.

## 10.5.4. Элементы простого анализа временных и вариационных рядов

В данном разделе представлен программный модуль **SimpleStat,** который экспортирует ряд полезных средств, предназначенных для обеспечения простого статистического анализа, прежде всего, для анализа вариационных и временных (*динамических*) рядов. *Экспортируемые* этим модулем средства предназначены для поддержки простого статистического анализа.

**SimpleStat –** ***поддержка функций простого статистического анализа***

Форматы вызова функций, экспортируемых модулем:

**SimpleStat:- X(*args*)** либо **with(SimpleStat): X(*args*)**

Формальные аргументы вызовов функций, экспортируемых модулем:

**X** – имя функции, экспортируемой **SimpleStat**-модулем
***args*** – формальные аргументы, соответствующие экспортируемым функциям

Описание программного модуля:

Простой модуль **SimpleStat** экспортирует ряд полезных функций, предназначенных для обеспечения простого статистического анализа, прежде всего, для анализа вариационных и временных (*динамических*) рядов. Модуль **SimpleStat** экспортирует тринадцать функций для поддержки простого статистического анализа данных, а именно:

***ACC, MCC, PCC, LT, FD, MAM, CR, CC, Sko, Ds, SR, LRM_NRM, Weights***

**SR(L)** – *средняя* статистических данных, заданных *списком*, *вектором* или текстовым *файлом* **L;**

**Ds(L)** – *дисперсия* статистических данных, заданных *списком*, *вектором* или текстовым *файлом*;

**Sko(L)** – *стандартное квадратичное отклонение* статистических данных, заданных списком, вектором или текстовым файлом данных **L;**

**CC(L1, L2** {**, t**}**)** – *коэффициент корреляции* между двумя множествами статистических данных, заданных списками, векторами либо текстовыми файлами **L1** и **L2; t –** (*необязательный*) положительное целое число, определяющее точность возвращаемого результата; по умолчанию полагается **t = Digits;** в дальнейшем **t**-аргумент имеет аналогичное назначение**;**

**CR(L1, L2 {, t}) –** *корреляционное отношение* между двумя множествами статистических данных, заданных списками, векторами или текстовыми файлами данных **L1** и **L2;**

**MAM(L, n) –** *скользящая средняя* уровней временного ряда, определенных списком, вектором либо текстовым файлом **L,** и **n –** длина скользящего интервала**;**

**FD(L) –** последовательность списков *разностей* всех порядков уровней временного ряда, определенного списком, вектором или текстовым файлом данных **L;**

**LT(L, F, X, t) –** служит для вычисления *тренда* временного ряда, заданного списком, вектором или текстовым файлом **L** значений его *уровней*. Другие формальные аргументы имеют следующий смысл: **F –** имя возвращаемой функции, определяющей искомый *линейный тренд* временного ряда; **X –** имя ведущей переменной функции **F** *линейного тренда*. Процедура **LT** не только возвращает уравнение, описывающее функцию *линейного тренда*, но и вычисляет определение данной линейной функции, обеспечивая в текущем сеансе доступ к ней на уровне средств главной *Maple*-библиотеки пакета**;**

**ACC(L) –** *коэффициент автокорреляции* статистических данных, заданных списком, вектором или текстовым файлом данных **L;**

**Weights(L) –** *веса* значений данных, заданных списком, вектором или текстовым файлом **L;** результат возвращается в виде (**2xn**) массива, чья первая строка представляет сами значения и вторая – соответствующие им веса;

**PCC(X, Y, Z) –** *частный коэффициент* корреляции между *тремя* множествами статистических данных, определенных списками, векторами или текстовыми файлами данных **X, Y, Z;**

**MCC(X, Y, Z) –** *множественный коэффициент* корреляции между тремя множествами статданных, определенных списками, векторами или текстовыми файлами **X, Y, Z;**

**LRM_NRM(A,B,*args*) –** *линейная/нелинейная модель регрессии* между двумя множествами статистических данных, заданных списками, векторами или текстовыми файлами **A** и **B;** описание формальных аргументов ***args*** может быть найдено в разделе **10.5.2** выше**.**

Данные в файлах должны быть значениями типов {*integer*, *float*, *fraction*}, разделенных символами перевода строки и возврата каретки {***Enter*** *клавиша***;** hex(0D0A)}. Более того, если длина одной из выборок меньше чем *три*, то возникает ошибочная ситуация, тогда как при обнаружении выборок с *разными* длинами производится их выравнивание по длине.

```
SimpleStat := module ()
local  n, AVZ63, L, G, SUM, Gdnt;
export      SR,          # средняя статистических данных
            Ds,          # дисперсия статистических данных
            Sko,         # стандартное квадратичное отклонение статистических данных
            CC,          # коэффициент корреляции
            CR,          # корреляционное отношение
            MAM,         # скользящая средняя уровней временного ряда
            FD,          # последовательность разностей всех уровней временного ряда
            LT,          # линейный тренд уровней временного ряда
            PCC,         # частный коэффициент корреляции
            MCC,         # множественный коэффициент корреляции
            ACC,         # коэффициент автокорреляции
```

```
            Weights,        # веса значений статистических данных
            LRM_NRM;        # линейная/нелинейная модель регрессии
description             "Simple statistical data analysis with Maple of releases 6, 7, 8, 9 and 10";
options                 `CopyRight (c) RANS_IAN = Tallinn-Grodno-Vilnius-Moscow; 30.08.2002`,
                        load = [AVZ63, SUM], package;
SR := proc(L::{list, vector, file})
local  a;
   `if`(type(L, 'list'), assign(a = L), `if`(type(L, 'vector'),  assign(a = convert(L, 'list1')),
    assign(a = map(op, convert(readdata1(L, 1), 'list1'))))); `+`(op(a))/nops(a)
end proc;
Ds := proc(L::{list, vector, file}) local  a, k;
    `if`(type(L, 'list'), assign(a=L), `if`(type(L, 'vector'), assign(a =convert(L, 'list1')),
     assign(a= map(op, convert(readdata1(L, 1), 'list1'))));
     sum((a[k] – SR(a))^2, k=1 .. nops(a))/nops(a)
end proc;
Sko := proc(L::{list, file}) local  a;
   `if`(type(L, 'list'), assign(a=L), assign(a = map(op, convert(readdata1(L, 1), 'list1')))); sqrt(Ds(a))
end proc;
CC := proc(A::{list, vector, file}, B::{list, vector, file})
local  a, b, o;
    `if`(type(A, 'list'), assign(a = A), `if`(type(A, 'vector'), assign(a = convert(A, 'list1')),
        assign(a = map(op, convert(readdata1(A, 1), 'list1')))); `if`(type(B, 'list'), assign(b = B),
       `if`(type(B, 'vector'), assign(b = convert(B, 'list1')),
       assign(b = map(op, convert(readdata1(B, 1), 'list1'))));
    assign(o = {nops(a), nops(b)}); `if`(o[1] < 3, ERROR("One of samples is too small"),
       [assign('a' = a[1 .. o[1]], 'b' = b[1 .. o[1]]), `if`(nops(o) > 1, WARNING("Lengths of samples is
       different, a levelling by minimal length has been done"), NULL)]);
    evalf((SR(AVZ63(a, b)) – SR(a)*SR(b))/sqrt(Ds(a)*Ds(b)), `if`(nargs = 3  and  type(args[3],
       'posint'), args[3], Digits))
end proc;
MAM := proc(L::{list, vector, file}, n::posint)
local  a, k, p, G;
    `if`(type(L, 'list'), assign(a = L), `if`(type(L, 'vector'), assign(a = convert(L, 'list1')),
        assign(a = map(op, convert(readdata1(L, 1), 'list1'))));
     [assign('G' = []), seq(assign('G' = [op(G), sum(a[p],
        p = k .. k + n – 1)/n]), k=1 .. nops(a) – n + 1), op(G)]
end proc;
LT := proc(L::{list, vector, file}, F::symbol, X::symbol, t::posint)
local  a, k, n, b;
    `if`(type(L, 'list'), assign(b = L), `if`(type(L, 'vector'), assign(b = convert(L, 'list1')),
```

```
        assign(b = map(op, convert(readdata1(L, 1), 'list1'))))); op([assign(a = [assign(n = nops(b)),
        assign(('%s') = 6*sum((2*k – n – 1)*b[k], k = 1 .. n)/(n^3 – n)), assign(('%g') = sum(b[k],
      k = 1 .. n)/n), %s, %g – 1/2*(n + 1)*%s, unassign('%s', '%g')]), RETURN(F(X)=evalf(a(b), t)[1]*
      X + evalf(a(b), t)[2], assign(F = proc(X)  options operator, arrow; evalf(a(b), t)[1]*X +
        evalf(a(b), t)[2]  end proc))])
end proc;
FD := proc(L::{list, vector, file})
local  k, p, FD1, A, K, b;
   `if`(type(L, 'list'), assign(b = L), `if`(type(L, 'vector'), assign(b = convert(L, 'list1')),
      assign(b = map(op, convert(readdata1(L, 1), 'list1')))));
    FD1 := (K)  ->  [assign('A' = []), seq(assign('A' = [op(A), K[p + 1] – K[p]]), p=1 .. nops(K) – 1),
      op(A)];   b, FD1(b), seq(FD1(A), k=2 .. nops(b) – 1)
end proc;
LRM_NRM := proc(A::{vector, file, list}, B::{vector, file ,list}, T::name, P::evaln, CR::evaln)
local  a, b, k, m, n, p, L, M, N, r, omega, o, G, X, Y, P1, P2, F, sr, ds, v;
  `if`(type(A, 'list'), assign(a = A), `if`(type(A, 'vector'), assign(a = convert(A, 'list1')),
      assign(a = map(op, convert(readdata1(A, 1), 'list1')))));
  `if`(type(B, 'list'), assign(b = B), `if`(type(B, 'vector'), assign(b = convert(B, 'list1')),
      assign(b = map(op, convert(readdata1(B, 1), 'list1')))));
  assign(o = {nops(b), nops(a)}), `if`(o[1] < 3, ERROR("One of samples is too small"),
      [assign('a' = a[1 .. o[1]], 'b' = b[1 .. o[1]]), `if`(1 < nops(o), WARNING("Lengths of
      samples is different, a levelling by minimal length has been done"), NULL)]);
  `if`(nops(a) <> nops(b), ERROR("different dimensionalities of lists of source statistical
      data"), assign(omega = [seq(1, p = 1 .. nops(a))], G = [], F = ['TIMES', 'BOLD', 10]));
  assign(v = (() ->  op([assign('L' = []), seq(assign('L' = [op(L), product(args[m][p],
      m = 1 .. nargs)]), p = 1 .. nops(args[1])), L])));
  assign(sr = (() ->  `+`(args)/nargs), ds = (() ->  `+`(seq((args[k] – sr(args))^2,
      k = 1 .. nargs))/nargs)), `if`(T = 'LRM', assign('M' = Matrix(2, 2, [[sr(op(v(b, b))),
      sr(op(b))], [sr(op(b)), 1]]), 'N' = Vector(2, [sr(op(v(a, b))), sr(op(a))])), `if`(T = 'NRM',
      assign('M' = Matrix(3, 3, [[sr(op(v(b, b, b, b))), sr(op(v(b, b, b))), sr(op(v(b, b)))],
      [sr(op(v(b, b, b))), sr(op(v(b, b))), sr(op(b))], [sr(op(v(b, b))), sr(op(b)), 1]]),
      'N' = Vector(3, [sr(op(v(a, b, b))), sr(op(v(a, b))), sr(op(a))])),
      ERROR("regression model type <%1> is inadmissible", T)));
  assign('r' = convert(evalf(LinearAlgebra:- LinearSolve(M, N)), 'list1'));
  assign(Y = `if`(nops(r) <> 1, proc(X)  local n; sum(r[n]*X^(nops(r) – n), n = 1 .. nops(r))  end proc,
      proc(X)  sign(op(r))*X  end proc)), `if`(nops(r) = 1, assign('CR' = sign(op(r))), NULL);
  `if`(nops(r) = 2, assign('CR' = evalf(sqrt((ds(op(a)) – sr(op(v(a – r[1]*b – r[2]*omega,
      a – r[1]*b – r[2]*omega))))/ds(op(a))))), `if`(nops(r) = 3, assign('CR' = evalf(sqrt((ds(op(a)) –
      sr(op(v(a – r[1]*v(b, b) – r[2]*b – r[3]*omega, a – r[1]*v(b, b) –
      r[2]*b – r[3]*omega))))/ds(op(a))))), NULL));
```

```
   seq(assign('G' = [op(G), [b[n], a[n]]]), n = 1 .. nops(a));
   P1 := plots['pointplot'](G, 'color' = 'blue', 'thickness' = 3, 'symbol' = 'CIRCLE', 'symbolsize' = 16);
   P2 := plot(Y(X), X = b[1] .. b[-1], 'thickness' = 2, 'color' = 'green');
   assign(P = plots['display']([P1, P2], 'axesfont' = F, 'scaling' = 'UNCONSTRAINED', 'labels' =
      ["B, X", "A, Y"], 'labeldirections' = ['HORIZONTAL', 'VERTICAL'], 'labelfont' = F)), sort(Y(X))
end proc;
CR := proc(A::{list,vector, file}, B::{list, vector, file})
local  a, b, p, cr;
   `if`(type(A, 'list'), assign(a = A), `if`(type(A, 'vector'), assign(a = convert(A, 'list1')),
      assign(a = map(op, convert(readdata1(A, 1), 'list1')))));
   `if`(type(B, 'list'), assign(b = B), `if`(type(B, 'vector') ,assign(b = convert(B, 'list1')),
      assign(b = map(op, convert(readdata1(B, 1), 'list1')))));  LRM_NRM(a, b, LRM, p, cr);
   evalf(cr, `if`(nargs = 3  and  type(args[3], 'posint'), args[3], Digits))
end proc;
PCC := proc(Z::{list, vector, file}, X::{list, vector, file}, Y::{list, vector, file})
local  a, b, c, o;
   `if`(type(Z, 'list'), assign(a = Z), `if`(type(Z, 'vector'), assign(a = convert(Z, 'list1')),
      assign(a = map(op, convert(readdata1(Z, 1), 'list1')))));
   `if`(type(X, 'list'), assign(b = X), `if`(type(X, 'vector'), assign(b = convert(X, 'list1')),
      assign(b = map(op, convert(readdata1(X, 1), 'list1')))));
   `if`(type(Y, 'list'), assign( c= Y), `if`(type(Y, 'vector'), assign(c = convert(Y, 'list1')),
      assign(c = map(op, convert(readdata1(Y, 1), 'list1')))));
   assign(o = {nops(a), nops(b), nops(c)}), `if`(o[1] < 3, ERROR("One of samples is too small"),
      [assign('a' = a[1 .. o[1]], 'b' = b[1 .. o[1]], 'c' = c[1 .. o[1]]),
      `if`(nops(o) > 1, WARNING("Lengths of samples is different, a levelling by minimal
      length has been done"), NULL)]);
   evalf((CR(a, b) – CR(a, c)*CR(b, c))/sqrt((1 – CR(a, c)^2*(1 – CR(b, c)^2))))
end proc;
MCC := proc(Z::{list, vector, file}, X::{list, vector, file}, Y::{list, vector, file})
local  a, b, c, k, delta, E, B, R, o;
   `if`(type(Z, 'list'), assign(a = Z), `if`(type(Z, 'vector'), assign(a = convert(Z, 'list1')),
      assign(a = map(op, convert(readdata1(Z, 1), 'list1')))));
   `if`(type(X, 'list'), assign(b = X), `if`(type(X, 'vector'), assign(b = convert(X, 'list1')),
      assign(b = map(op, convert(readdata1(X, 1), 'list1')))));
   `if`(type(Y, 'list'), assign( c= Y), `if`(type(Y, 'vector'), assign(c = convert(Y, 'list1')),
      assign(c = map(op, convert(readdata1(Y, 1), 'list1')))));
   assign(o = {nops(a), nops(b), nops(c)}), `if`(o[1] < 3, ERROR("One of samples is too small"),
      [assign('a' = a[1 .. o[1]], 'b' = b[1 .. o[1]], 'c' = c[1 .. o[1]]), `if`(nops(o) > 1,
      WARNING("Lengths of samples is different, a levelling by minimal length has been
      done"), NULL)]);
```

```
    E := []:  for k to nops(a)  do  E:= [op(E), 1]  end do:  k:= 'k':  delta := matrix(3, 3, [SUM(b, b),
        SUM(b, c), SUM(b, E), SUM(b, c), SUM(c, c), SUM(c, E), SUM(b, E), SUM(c, E), nops(a)]);
    B:= vector(3, [SUM(a, b), SUM(a, c), SUM(a, E)]);  R := evalf(linalg['linsolve'](delta, B));
    evalf(sqrt((R[1]*PCC(a, b, c)*sqrt(Ds(b)) + R[2]*PCC(a, c, b)*sqrt(Ds(c)))/sqrt(Ds(a))))
end proc;
AVZ63 := proc()
local  k, p, L;
   op([assign(L = []), seq(assign('L' = [op(L), product(args[k][p], k=1 .. nargs)]),
      p=1 .. nops(args[1])), L])
end proc;
SUM := proc(A::list, B::list)  local k;  sum(A[k]*B[k], k=1 .. nops(A))  end proc;
module Gdnt ()
local  p, h, n, Sr, Ds, S;
export  ACC;
description "Calculation of autocorrelation coefficient";
   Sr := (S, p)  ->  sum(S[k], k = p .. `if`(p = 1, nops(S) – 1, nops(S)))/(nops(S) – 1);
   Ds := (S, h)  ->  sum(S[k]^2, k = h .. `if`(h = 1, nops(S) – 1, nops(S)))/(nops(S) – 1) – Sr(S, h)^2;
   ACC := (S)  ->  evalf(sum((S[n + 1] – Sr(S, 2))*(S[n] – Sr(S, 2)), n=1 .. (nops(S) – 1))/
            (sqrt(Ds(S, 1)*Ds(S, 2))*(nops(S) – 1)))
end module;
ACC := proc(L::{list, vector, file})
local  a;
   `if`(type(L, 'list'), assign(a = L), `if`(type(L, 'vector'), assign(a = convert(L, 'list1')),
        assign(a = map(op, convert(readdata1(L, 1), 'list1')))));   Gdnt:- ACC(a)
end proc;
Weights := proc(L::{list, vector, file})
local  a, k, p, G, R;
   `if`(type(L, 'list'), assign(a = L), `if`(type(L, 'vector'), assign(a = convert(L, 'list1')),
        assign(a = map(op, convert(readdata1(L, 1), 'list1')))));  G := [op(sort({op(a)}))]:
        R := array('sparse', 1 .. 2, 1 .. nops(G)):
   for k to nops(G) do  R[1, k] := G[k]:   for p to nops(a)  do
   if  G[k] = a[p]  then  R[2, k] := R[2, k] + 1  else next  end if
      end do
     end do;
     evalm(R)
   end proc;
end module
```

Типичные примеры применения средств, экспортируемых программным модулем:

```
> with(SimpleStat);  F:= "C:\\Academy\\Examples\\Sample2.dat":
```

[*ACC*, *CC*, *CR*, *Ds*, *FD*, *LRM_NRM*, *LT*, *MAM*, *MCC*, *PCC*, *SR*, *Sko*, *Weights*]

**> SR(F), Ds(F), Sko(F);** ⇒ 81.30769231, 137.2899410, 11.71707903

**> CC("C:\\Academy\\Examples\\Members.dat", "C:\\Academy\\Examples\\Age.dat");**

-0.9830255251

**> MAM(F, 3); LT(F, H, X, 5); FD(F);**

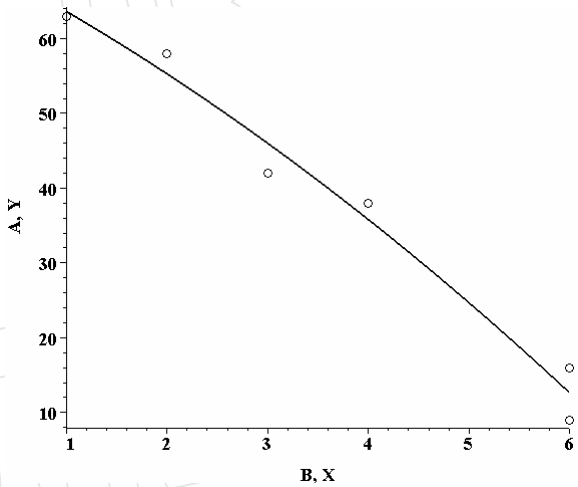
[73.66666667, 74.66666667, 81.33333333, 89.33333333, 88.33333333, 89.66666667,
86.00000000, 78.00000000, 70.00000000, 62.00000000, 62.00000000, 65.66666667,
69.33333333, 73.00000000, 78.33333333, 83.66666667, 89.00000000, 89.00000000,
89.00000000, 86.66666667, 84.33333333, 87.00000000, 92.00000000, 97.00000000]

H(X) = 0.58735*X + 73.378

[86., 62., 73., 89., 82., 97., 86., 86., 86., 62., 62., 62., 62., 73., 73., 73., 89., 89., 89.,
89., 89., 82., 82., 97., 97., 97.], [-24., 11., 16., -7., 15., -11., 0., 0., -24., 0., 0., 0., 11., 0.,
0., 16., 0., 0., 0., 0., -7., 0., 15., 0., 0.], [35., 5., -23., 22., -26., 11., 0., -24., 24., 0., 0., 11.,
-11., 0., 16., -16., 0., 0., 0., -7., 7., 15., -15., 0.], [ -30., -28., 45., -48., 37., -11., -24., 48.,
178., -133., 35., 85., -144., 96., -13., -44., 66., -28., -53., 96., -64., 16., -7., 28., 83.], ….]

**> LRM_NRM("c:\\academy\\examples\\Members.dat", "c:\\academy\\examples\\Age.dat", NRM, a, b), b; a;**

6.394999349 - 0.0001390633502*X^2 - 0.07160952800*X, 0.9516896937




**> CR("C:/Academy/Examples/Members.dat", "C:/Academy/Examples/Age.dat");** ⇒ 0.98302553

**> PCC("C:\\Academy\\Examples\\Members.dat", "C:\\Academy\\Examples\\Age.dat", F);**

-0.7446177489

**> MCC("C:\\Academy\\Examples\\Members.dat","C:\\Academy\\Examples\\Age.dat", F);**

0.8548484302

**> ACC("C:\\academy/examples\\age.dat"), Weights("D:\\academy/examples/sample2.dat");**

$$0.9194529096, \begin{bmatrix} 86. & 62. & 73. & 89. & 82. & 97. & 61. & 56. & 7. & 87. \\ 6 & 7 & 11 & 6 & 1 & 3 & 6 & 1 & 1 & 1 \end{bmatrix}$$

Средства, представленные в данной главе, могут служить в качестве *дополнения* к средствам пакета, обеспечивая решение достаточно широкого круга задач, имеющих дело с простой *обработкой* статистических данных. С учетом представленных выше рекомендаций данные средства могут быть существенно расширены в функциональном отношении, одновременно предоставляя хорошую основу для получения практического навыка программирования в среде пакета *Maple* задач статистического анализа данных.

# Литература


1. **Aladjev V.Z.** *To Theory of Homogeneous Structures*.- Tallinn: Estonian Academic Press, 1972, 235 pp. (*in Russian with extended English summary*)
2. **Aladjev V.Z.** *Introduction to Operating System of **EC**-computers*.- Tallinn, 1975, 156 pp. (*in Russian*)
3. **Aladjev V.Z.** *Introduction to Architecture of **EC**-computers*.- Tallinn: Valgus Press, 1976, 332 pp. (*in Russian with extended English summary*)
4. **Aladjev V.Z.** *Mathematical Theory of Homogeneous Structures and Their Applications*.- Tallinn: Valgus Press, 1980
5. **Aladjev V.Z. et al.** *Mathematical Biology of Development*.- Moscow: Nauka Press, 1982, 254 pp. (*in Russian with extended English summary*)
6. **Aladjev V.Z.** *The Architecture and Software of **CM**-computers*.- Tallinn: IVC Gossbanka USSR, 1983 (*in Russian with extended English summary*)
7. **Aladjev V.Z.** *Course of Lectures on Personal Computer **ISKRA-226.-*** Tallinn: SKB MPSM, 1986 (*in Russian with extended English summary*)
8. **Aladjev V.Z. et al.** *Personal Computer **ISKRA-226.-*** Kiev: Ukrainian Encyclopedia Press, 1988, 150 pp. (*in Russian with extended English summary*)
9. **Aladjev V.Z.** *Software Engineering for PC **ISKRA-226.-*** Kiev: Technics Press, 1989, 255 pp. (*in Russian with extended English summary*)
10. **Aladjev V.Z., Sirodza I.B.** *Solution of Engineering Problems in Basic-language of PC **ISKRA-226.-*** Harkov: Harkov Aircraft Institute, 1988, 128 pp. (*in Russian with extended English summary*)
11. **Aladjev V.Z., Shilenko V.F.** *Personal Computer **ISKRA-1030.-*** Kiev: Ukrainian Soviet Encyclopedia Press, 1990 (*in Russian with extended English summary*)
12. **Aladjev V.Z.** *Homogeneous Structures: Theoretical and Applied Aspects*.- Kiev: Technics Press, 1990 (*in Russian with extended English summary*)
13. **Aladjev V.Z., Sirodza I.B.** *Personal Computer **ISKRA-1030:** Toolkit and Designing of Programs*.- Moscow: Higher School Press, 1991 (*in Russian with extended English summary*)
14. **Aladjev V.Z.** *Scientific and Technical CAD*.- Kiev: Ukrainian Soviet Encyclopedia Press, 1991 (*in Russian with extended English summary*)
15. **Aladjev V.Z., Gershgorn N.A.** *Computational Problems on Personal Computer*.- Kiev: Technics Press, 1991 (*in Russian with extended English summary*)
16. **Aladjev V.Z. et al.** *Utilities for Personal Computer*.- Kiev: Technics Press, 1991 (*in Russian*)
17. **Aladjev V.Z., Sirodza I.B.** *Computer Mixture*.- Tallinn: CART Press, 1992 (*in Russian*)
18. **Aladjev V.Z., Tupalo V.G.** *Computer Reader*.- Kiev: Ukrainian Soviet Encyclopedia Press, 1993 (*in Russian with extended English summary*)
19. **Aladjev V.Z., Tupalo V.G.** *Turbo-Pascal for All*.- Kiev: Technics Press, 1993, ISBN 5-335-01225-0 (*in Russian with extended English summary*)
20. **Aladjev V.Z., Tupalo V.G.** *Scientific Editing and Statistics on Personal Computer*.- Moscow: Mintopenergo Press, 1993 (*in Russian with extended English summary*)
21. **Aladjev V.Z., Tupalo V.G.** *Computer Telecommunication*.- Moscow: Mintopenergo Press, 1993 (*in Russian with extended English summary*)

**22. Aladjev V.Z., Tupalo V.G.** *Algebraic Computations on Computer*.- Moscow: Mintopenergo Press, 1993 (*in Russian with extended English summary*)

**23. Aladjev V.Z., Tupalo V.G.** *Scientific and Practical Activity of the TRG during 1969-1993*.- Moscow: Mintopenergo Press, 1994 (*in Russian with extended English summary*)

**24.** *Software of **EC**-computers and Management Information Systems* **/** Ed. **V.Z. Aladjev.-** Tallinn: Valgus Press, 1978 (*in Russian with extended English summary*)

**25.** *Database Management System on the Basis of Operating System **MINIOS** and DBMS **OKA*** **/**Ed. **V.Z. Aladjev.-** Tallinn: Valgus Press, 1980 (*in Russian with extended English summary*)

**26.** *Parallel Systems of Information Processing and Parallel Algorithms***/**Ed. **V.Z. Aladjev.-** Tallinn: Valgus Press, 1981 (*in Russian with extended English summary*)

**27.** *Parallel Systems of Information Processing***/**Ed. **V.Z. Aladjev.-** Tallinn: Valgus Press, 1983

**28.** *Structural-Analytical Models and Algorithms of Recognition and Identification of Control Objects* **/**Ed. **V.Z. Aladjev.-** Kiev: Technics Press, 1993 (*in Russian with extended English summary*)

**29. Aladjev V.Z., Veetyusme R.A., Hunt U.J.** *General Theory of Statistics*.- Tallinn: TRG & Salcombe Eesti Ltd., 1995, ISBN 1-995-14642-8 (*in Russian with extended English summary*)

**30. Veetyusme R.** *Information needs in democratic market economy*, in [32]

**31.** *Statistical Theory and Methods* **//** Abstracts.- T.& A. CONSTABLE Ltd., Edinburgh, 1995

**32.** *Statistical Information Systems in a Market Economy*, Paris: EUROSTAT, 1993

**33. Astola J.** *A Measure of Overal Statistical Dependence Based on the Entropy Concept*. -Vaasa: University of Vaasa, 1983

**34.** *The **SAS** System*.- Cary: SAS Institute Inc., 1994, 2195 pp.

**35. Berenson M.L. et al.** *Intermediate Statistical Methods and Applications: A Computer Package Approach*.- New Jersey: Prentice-Hall, 1983

**36. Brockwell P., Davis R.** ***ITSM*** *for Windows: User Guide to Time Series Modeling and Forecasting*.- Berlin: Springer-Verlag, 1994

**37. Diday E.** *New Approaches in Classification and Data Analysis*.- Berlin: Springer, 1994

**38. Batz W.** *Das **SAS** Survival-Handbuch*.- Oslo: Springer, 1994

**39. Janssen J.P.** *Statistische Datenanalyse mit **SPSS** fur Windows*.- Berlin: Springer, 1994 (*in German*)

**40. Lancaster H.** *Quantitative Methods in Biological and Medical Sciences*.- Berlin: Springer, 1994

**41. Malley J.** *Statistical Applications of Jordan Algebras*.- Berlin: Springer-Verlag, 1994

**42. Polasek W.** ***EDA -*** *Explorative Datenanalyse*.- Berlin: Springer-Verlag, 1994

**43. Roe B.P.** *Probability and Statistics in Experimental Physics*.- Springer-Verlag, 1994

**44. Tanner M.A.** *Tools for Statistical Inference*.- Berlin: Springer-Verlag, 1994

**45. Venables W.N., Riley B.D.** *Modern Applied Statistics with **S-Plus**.-* Berlin: Springer-Verlag, 1998

**46. Weerahandi S.** *Exact Statistical Methods in Data Analysis*.- Heidelberg: Springer-Verlag, 1994.

**47. Daniel C., Wood F.S.** *Computer Analysis of Multifacto Data*.- N. Y.: John Wiley& Sons, 1980

**48. Box G.E. et al.** *Statistics for Experimenters*.- N.Y.: John Wiley& Sons, 1978

**49. Afifi A.A., Azen S.** *Statistical Analysis*.- N.Y.: Academic Press, 1979

**50. Dixon W.J. et al.** ***BMDP*** *Biomedical Computer Programs*.- Berkeley, 1981

**51. Merkov A.M.** *General Theory and Technique of Sanitary: Statistical Researches*.- Moscow: Medgiz, 1960 (*in Russian with extended English summary*)

**52. Beili N.** *Statistical Methods in Biology*.- Moscow: Mir Press, 1963 (*in Russian*)

**53. Urbah V.J.** *Biometric Methods*.- Moscow: Nauka Press, 1964 (*in Russian*)

**54. Bessmertny B.S.** *Mathematical Statistics in Clinical, Preventive and Experimental Medicine*.- Moscow: Medicine Press, 1967 (*in Russian with extended English summary*)

**55. Koroljuk V.S. et al.** *The Reference Book on Probability Theory and Mathematical Statistics*.- Moscow: Nauka Press, 1985 (*in Russian with extended English summary*)
**56. Krenkel T.** *Personal Computers in Engineering Practice*.- Moscow: Energy Press, 1989 (*in Russian*).
**57. Kozlov M.V.** *Introduction to Mathematical Statistics*.- Moscow: Moscow State University, 1987 (*in Russian with extended English summary*)
**58. Pugacshev V.S.** *Probability Theory and Mathematical Statistics*.- Moscow: Nauka Press, 1979 (*in Russian with extended English summary*)
**59. Fihtengoltz G.** *Course of Differential and Integral Calculus*.- Moscow: Fizmatgiz, 1960 (*in Russian*)
**60. Bikel P.** *Mathematical Statistics*.- Moscow: Finance and Statistics Press, 1983 (*in Russian*)
**61. Smirnov N.V., Dunin-Barkovsky I.V.** *Course of Probability Theory and Mathematical Statistics*.- Moscow: Nauka Press, 1969 (*in Russian with extended English summary*)
**62. Ventzel E.S.** *Applied Problems of Probability Theory*.- Moscow: Radio and Communication, 1983 (*in Russian with extended English summary*)
**63. Ventzel E.S.** *Probability Theory*.- Moscow: Nauka Press, 1964 (*in Russian*).
**64. Rjabushkin T.V.** *General Theory of Statistics*.- Moscow: Finance and Statistics Press, 1981 (*in Russian with extended English summary*)
**65. Bojarsky A.J. et al.** *General Theory of Statistics*.- Moscow: Moscow State University, 1977 (*in Russian with extended English summary*)
**66. Drucshinin N.K.** *Mathematical Statistics in Economics*.- Moscow: Statistics Press, 1971 (*in Russian*)
**67. Donda A. et al.** *Statistics*.- Moscow: Statistics Press, 1974.
**68. Dolgushevsky V.L.** *Collected Problems on General Theory of Statistics*.- Moscow: Statistics Press, 1966 (*in Russian with extended English summary*)
**69. Baideldinov L.A.** *Statistics in Sociological Researches*.- Alma-Ata: Nauka Press, 1965 (*in Russian*)
**70. Kimbel G.** *How Correctly to Use Statistics*.- Moscow: Statistics Press, 1982 (*in Russian*)
**71. Egermaer F et al.** *Fundamentals of Statistics*.- Moscow: Gosstatizdat Press, 1961 (*in Russian*)
**72. Zamoskovny O.P., Hazanov J.G.** *Fundamentals of Statistics*.- Moscow: Statistics Press, 1974 (*in Russian with extended English summary*)
**73. Grankov V.P.** *Sample Observation*.- Moscow: Gosstatizdat Press, 1963 (*in Russian*)
**74. Erenburg A.G.** *Analysis and Interpretation of Statistical Data*.- Moscow: Finance Press, 1981 (*in Russian with extended English summary*)
**75. Maslov P.P.** *Statistics*.- Moscow: Mysl Press, 1964 (*in Russian with extended English summary*)
**76. Gerschuk J.P.** *Charts in Mathematical and Statistical Analysis*.- Moscow: Statistics Press, 1972 (*in Russian with extended English summary*)
**77. Broudi M.B.** *About Statistical Reasoning*.- Moscow: Statistics Press, 1968 (*in Russian*)
**78. Meiesaar K.I.** *Statistical Observation*.- Tartu: Tartu State University, 1985 (*in Russian*)
**79. Harlamov A.I.** *Absolute and Relative Values*.- Moscow: VZFEI Press, 1959 (*in Russian*)
**80. Jinni K.** *The Logic in Statistics*.- Moscow: Statistics, 1973 (*in Russian with English summary*)
**81. Daitbegov D.G.** *The Software of Statistical Data Processing*.- Moscow, 1984 (*in Russian*)
**82. Ljalin V.S. et al.** *Statistics*.- Moscow: Mysl Press, 1985 (*in Russian with English summary*)
**83. Kun J.** *Descriptive and Inductive Statistics*.- Moscow: Finance and Statistics Press, 1981 (*in Russian*)
**84. Ploshko B.G.** *History of Statistics*.- Moscow: Finance and Statistics Press, 1990 (*in Russian*)
**85. Jessen R.** *Methods of Statistical Examinations*.- Moscow: Finance and Statistics, 1985 (*in Russian*)
**86. Gozulov A.I.** *Collected Problems on Statistics*.- Moscow: Statistics Press, 1969 (*in Russian*)
**87. Vainberg J., Shumeker J.** *Statistics*.- Moscow: Statistics Press, 1979 (*in Russian*)

**88. Nikitina E.D.** *Collection of Definitions of the Term "Statistics"*.- Moscow: Moscow State University, 1972 (*in Russian with extended English summary*)
**89. Gren E.** *Statistical Games and Their Application*.- Moscow: Statistics Press, 1975 (*in Russian*)
**90. Maslov P.P.** *Computing Technique with Digits*.- Moscow: Statistics Press, 1977 (*in Russian*)
**91. Gaskarov D.V., Shapovalov V.** *Small Sampling*.- Moscow: Statistics Press, 1978 (*in Russian*)
**92. Kildishev G.S.** *General Theory of Statistics*.- Moscow: Statistics Press, 1980 (*in Russian*)
**93. Duait G.B.** *Integral Tables and Other Mathematical Formulas*.- Moscow: Nauka Press, 1973 (*in Russian with extended English summary*)
**94. Korn G.** *The Reference Book on Mathematics for Researchers*.- Moscow: Nauka Press, 1973 (*in Russian with extended English summary*)
**95.** *The Brief Economical Dictionary*.- Moscow: Politizdat Press, 1987 (*in Russian*)
**96.** *The Statistical Dictionary*.- Moscow: Statistics Press, 1990 (*in Russian with English summary*)
**97. Char B.W.** ***Maple V*** *Library Reference Manual*.- Berlin: Springer-Verlag, 1993
**98. Redfern D.** *The* ***Maple*** *Handbook*.- Berlin: Springer, 1994
**99. Burkhardt W.** *First Steps in* ***Matematica***.- Berlin: Springer- Verlag, 1994
**100.** ***SAS*** *Institute Publications*.- Cary: SAS Institute Inc., 1994
**101.** ***SAS/STAT*** *User`s Guide*.- Cary: SAS Institute Inc., 1990
**102.** ***STATISTICA*****.-** Tulsa: StatSoft Inc., 1994
**103.** *The Program on Course "General Theory of Statistics"*.- Minsk: Institute of Modern Knowledge, 1995 (*in Russian with extended English summary*)
**104. Shannon C.** *Activities on Information Theory and Cybernetics*.- Moscow: Mir Press, 1963
**105.** *Statistical Publications*.- Tallinn: Statistical Office of Estonia, 1995.
**106. Linnik J.V.** *Lectures on Problems of Analytical Statistics*.- Moscow: Nauka Press, 1994 (*in Russian*)
**107. Oschegov S.V.** *Explanatory Dictionary of Russian*.- Moscow: AZ Press, 1994 (*in Russian*)
**108. Dorogovcev A.Ju.** *Calculus*.- Kiev: Higher School Press, 1985 (*in Russian*)
**109. Lashko I.G.** *Reference Manual on Calculus*.- Kiev: Technics Press, 1986 (*in Russian*)
**110. Rozin B.T.** *The Theory of Pattern Recognition in Economical Researches*.- Moscow: Statistics Press, 1973 (*in Russian with extended English summary*)
**111. Fihtengolz G.M.** *Fundamentals of Calculus*, vol. **1-2.-** Moscow: GITTL Press, 1957 (*in Russian*)
**112. Suslov I.P.** *Fundamentals of the Theory of Veracity of Statistics*.- Novosibirsk: Nauka Press, 1979 (*in Russian with extended English summary*)
**113. Martynov V.** *International Statistics*.- Moscow: Statistics Press, 1974 (*in Russian*)
**114. Savinsky D. et al.** *General Theory of Statistics*.- Moscow: Moscow University, 1960 (*in Russian*)
**115.** *The Statistical Dictionary*.- Moscow: Finance and Statistics Press, 1989 (*in Russian*)
**116. Suslov I.P.** *General Theory of Statistics*.- Moscow: Statistics Press, 1970 (*in Russian*)
**117. Reihman U.Z.** *Application of Statistics*.- Moscow: Statistics Press, 1969 (*in Russian*)
**118. Minium E.W.** *Elements of Statistical Reasoning*.- N. Y.: John Wiley& Sons, 1982
**119.** *The Philosophical Dictionary*.- Moscow: Politizdat, 1980
**120.** *Mathematical Encyclopedia*, vol. **1.-** Moscow: Soviet Encyclopedia, 1977, pp. 50-53 (*in Russian*)
**121.** *Teabevihik*.- Tartu: Eesti Statistikaselts, 1994 (*in Estonian*)
**122. Kelley T.L.** *Fundamentals of Statistics*.- London: Harvard University Press, 1947
**123. Zelditch M.** *A Basic Course in Sociological Statistics*.- New York: Henry Holt and Company, 1959
**124. Goedicke V.** *Introduction to the Theory of Statistics*.- N.Y.: Harper & Brothers Publishers, 1953
**125. Connor L.R.** *Statistics in Theory and Practice*.- London: Pitman&Sons, 1934

**126. Rouanet H. et al.** *New Ways in Statistical Methodology*.- Paris, 1998
**127. Aladjev V.Z., Hunt U.Ja., Shishakov M.L.** *Course of General Theory of Statistics*.- Gomel: BELGUT Press, 1995 (*in Russian with extended English summary*)
**128. Hunt U., Shishakov M.** *Probability Theory and Mathematical Statistics* / Ed. acad. **V.Z. Aladjev.-** Gomel: Russian Academy of Cosmonautics, 1997 (*in Russian with extended English summary*)
**129. Tooming L.** StatistikaSõnastik. Eesti-Inglise-Saksa-Vene.- Tartu: Tartu University Press, 1996
**130. Nelson B.L.** *Elements of Modern Statistics: For Students of Economics and Business*.- N.-Y.: Appleton-Century, 1961
**131. Haber A., Runyon R.** *General Statistics*.- London: Addison-Wesley Publishing Co., 1969
**132. McPherson G.** *Statistics in Scientific Investigation: Its Basis, Application and Interpretation*.- Berlin: Springer-Verlag, 1990
**133. Leach C.** *Introduction to Statistics: A Nonparametric Approach for the Social Sciences*.- Toronto: John Wiley, 1979
**134. Aladjev V.Z., Hunt U.J., Shishakov M.L.** *Mathematics on Personal Computer*.- Gomel: BELGUT Press, 1996 (*in Russian with extended English summary*)
**135. Aladjev V.Z., Shishakov M.L.** *Introduction to Package* ***Mathematica*** **2.2.-** Moscow: FILIN Press, 1997 (*in Russian with extended English summary*)
**136. Aladjev V.Z., Hunt U.J., Shishakov M.L.** *Fundamentals of Computer Informatics*.- Gomel: TRG& Salcombe Eesti Ltd. & Russian Academy of Noosphere, 1997 (*in Russian with English summary*)
**137. Aladjev V.Z., Hunt U.J., Shishakov M.L.** *Basics o Informatics*.- Moscow: FILIN Press, 1998 (*in Russian with extended English summary*)
**138. Aladjev V.Z., Hunt U.Ja., Shishakov M.L.** *Basics of Informatics*. 2nd edition.- Moscow: FILIN Press, 1999 (*in Russian with extended English summary*)
**139. Aladjev V.Z., Vaganov V.A., Hunt U.J, Shishakov M.L.** *Introduction to Mathematical Package* ***Maple V.-*** Gomel: Russian Academy of Noosphere, 1998 (*in Russian with English summary*)
**140. Aladjev V.Z., Vaganov V.A., Hunt U.J, Shishakov M.L.** Programming in Mathematical Package ***Maple V.-*** Tallinn-Gomel-Moscow: TRG, 1999 (*in Russian with English summary*)
**141. Aladjev V.Z., Vaganov V.A., Hunt U.J., Shishakov M.L.** A Workstation for Mathematicians.- Tallinn-Gomel-Moscow: VASCO & Salcombe Eesti Ltd., 1999 (*in Russian with English summary*)
**142. Aladjev V.Z., Shishakov M.L., Trohova T.A.** *Basics of Computer Informatics*.- Minsk: Tetrasystems Press, 2000 (*in Russian with extended English summary*)
**143. Aladjev V.Z., Bogdevicius M.A.** *Solution of Physical, Technical and Mathematical Problems with* ***Maple V.*** Vilnius: VGTU Press, 1999 (*in Russian with extended English summary*)
**144. Aladjev V.Z., Shishakov M.L.** *A Workstation for Mathematicians*.- Moscow: BINOM Press, 2000 + CD (*in Russian with extended English summary*)
**145.** *Statistics Sources* **/** Eds. **J. O`Brien, S. Wasserman.-** London: Gale Research Inc., 1991
**146. Lenin V.I.** *Complete Works*, 5-th edition, vol. 1-55.- Moscow: Politizdat Press, 1967 (*in Russian*)
**147. Shurenkov V.** *The Ergodic Theory of Markov Processes*.- Moscow: Nauka Press, 1989 (*in Russian*)
**148. Ventzel A.D.** *Course of Theory of Stochastic Processes*.- Moscow: Nauka Press, 1990 (*in Russian*)
**149. Pugachev V., Sinicyn I.** *Stochastic Differential Systems*.- Moscow: Nauka Press, 1990 (*in Russian*)
**150. Polischuk L.I.** *Analysis of Multiple-way Economic and Mathematical Models*.- Novosibirsk: Nauka Press, 1989 (*in Russian with extended English summary*)
**151.** *Functionals of Stochastic Processes and Statistical Conclusions*.- Tashkent: FAN Press, 1989 (*in Russian with extended English summary*)
**152. Nummelin E.** *General Irreducible Markov Chains and Non-Negative Operators*.- London: Cambridge Press, 1984

**153. Aladjev V.Z.** To an Asymptotical Property of Stochastic Homogeneous Structures **//** Proc. AN ESSR. Fiz.-Math., **20,** no. **2**, 1971 (*in Russian with extended English summary*)
**154. Bogoljubov N.N. et al.** *Mathematical Methods of Statistical Mechanics of Model Systems*.- Moscow: Nauka, 1989 (*in Russian with extended English summary*)
**155. Lakin G.F.** *Biometrics*.- Moscow: Higher School, 1990 (*in Russian with extended English summary*)
**156. Kobayashi M.** ***Mathematica****: An Introduction to Statistics and Probability*.- Tokyo: Toppan, 1994
**157. Hildebrand D., Ott R.** *Statistical Thinking for Managers*, 4th Ed.- London: Brooks/Cole, 1998
**158. Karian Z., Tanis E.** *Probability and Statistics Explorations with* ***Maple***.- N.Y.: Prentice Hall, 1995
**159. Roe B.,** *Probability and Statistics in Experimental Physics*.- Berlin: Springer-Verlag, 1998
**160. Chase W.** *General Statistics*, 4th Edition.- N.Y.: John Wiley & Sons, 1999
**161. Kinney J.J.,** *Probability: An Introduction with Statistical Applications*.- N.Y.: John Wiley, 1996
**162. Kallenberg O.** *Foundations of Modern Probability*.- Oslo-Berlin: Springer-Verlag, 1997
**163. Lange K.** *Mathematical and Statistical Methods for Genetic Analysis*.- Oslo: Springer-Verlag, 1997
**164. Prabhu N.** *Stochastic Storage Processes*.- N.Y.: Springer-Verlag, 1998
**165. Voelki K.E., Gerber S.B.** *Using* ***SPSS*** *for Windows*.- N.Y.: Springer-Verlag, 1999
**166. Råde L., Westergren B.** *Mathematics Handbook*.- Berlin: Springer-Verlag, 1999
**167. Øksendal B.** *Stochastic Differential Equations*.- Oslo: Springer-Verlag, 1998, ISBN 3-540-63720-6
**168. Rosanov Y. A.** *Random Fields and Stochastic Partial Diiferential Equations*.- London: Kluwer, 1998
**169. Härdle W., Klinke S., Müller M.** ***XploRe*** **–** *Academic Edition: The Interactive Statistical Computing Environment*.- Berlin: Springer-Verlag, 2000 + CD, ISBN 3-540-14767-5
**170. Kloeden P.E., Platen E.F.** *Numerical Solution of Stochastic Differential Equations*.- Heidelberg: Springer-Verlag, 1999
**171. Beltrami E.** *What is Random: Chance and Order in Mathematics and Life*.- N.Y.: Springer, 1999
**172. Brzezniak Z., Zastawniak T.** *Basic Stochastic Processes*.- London: Springer-Verlag, 1999
**173. Karatzas I., Shreve S.** *Methods of Mathematical Finance*.- N.Y.: Springer-Verlag, 1999
**174. Serfozo R.** *Introduction to Stochastic Networks*.- N.Y.: Springer-Verlag, 1999, ISBN 0-387-98773-8
**175. Grimmett G.** *Percolation*.- London: Springer-Verlag, 1999
**176. Liggett T.** *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*.- N.Y.: Springer-Verlag, 1999
**177. Liptser R., Shiryaev A.** *Statistics of Random Processes*, vol. **1-2.-** Berlin: Springer-Verlag, 1999
**178. Jacod J., Proter P.** *Probability Essentials*.- Paris: Springer-Verlag, 1999, ISBN 3-540-66419-X
**179.** *Limit Theorems of Probability Theory* **/** Eds. **Prokhorov Y., Statulevicius V.-** Berlin: Springer, 1999
**180. Dale A.I.** *A History of Inverse Probability: From Thomas Bayes to Karl Pearson*.- London: Springer-Verlag, 1999
**181. Bertoin J. et al.** *Lectures on Probability Theory and Statistics*.- Paris: Springer-Verlag, 1999
**182. Mittelhammer R.C.** *Mathematical Statistics for Economics and Business*.- N.Y.: Springer, 1999
**183. Pitman J.** *Probability*.- Berlin: Springer-Verlag, 1999.
**184. Li X.R.** *Probability Random Signals and Statistics*.- Berlin: Springer-Verlag, 1999
**185. Karian Z., Dudewicz E.J.** *Modern Statistical System and* ***GPSS*** *Simulation*.- Berlin: Springer, 1998
**186. Harris J., Stocker H.** *Handbook of Mathematics and Computational Science*.- Oslo: Springer, 1998
**187. Harville D.A.** *Matrix Algebra from a Statistician's Perspective*.- Berlin: Springer-Verlag, 1998
**188.** *Encyclopedia of Statistical Sciences* / Eds. **S. Kotz and N. Johnson,** vol. **1-9.-** N.Y.: Wiley, 1995
**189. Kotz S.** *Russian-English/English-Russian Glossary of Statistical Terms*.- London: Oliver, 1971
**190. Aladjev V.Z. et al.** *The Electronic Library of the Books*.- Tallinn: Tallinn Research Group, 1999

**191. Aladjev V.Z., Hunt U.J., Shishakov M.L.** *Research Activity of the Tallinn Research Group: Scientific Report for Period* 1995-1998 **/** Ed. **A. Ursul.-** Tallinn: TRG Press, 1998

**192. Aladjev V.Z. et al.** The Tallinn Research Group and International Academy of Noosphere: Scientific and Applied Activity for Period 1995-1999, in [143], pp. 668-678.

**193. Gordon H.** *Classification: Methods for Exploratory Data Analysis*.- London: Chapman & Hall, 1981

**194. Schiffman S.S. et al.** *Introduction to Multivariate Scaling*.- N.Y.: Academic Press, 1981

**195. Chatfield C., Collins A.** *Introduction to Multivariate Analysis*.- London: Chapman and Hall, 1980

**196. Golovin B.N.** *Language and Statistics*.- Moscow: Nauka Press, 1971 (*in Russian*)

**197. Gorevaja V.S.** *Statistical Description of Function-style Subdivisions of Modern English*.- Kalinin, 1974 (*in Russian with extended English summary*)

**198.** *Probability-statistical Organization of Neural Mechanisms of the Brain* **/** Eds. **Kogan A. et al.-** Rostov, 1974 (*in Russian with extended English summary*)

**199.** *Estonian Labour Forse 1989-1999*.- Tallinn: Statistical Office of Estonia, 1999 (*in Estonian&English*)

**200.** *Statistical Yearbook of Estonia*.- Tallinn: Statistikaamet Press, 1999 (*in Estonian and English*)

**201.** *Statistical Yearbook of Russia*.- Moscow: Goskomstat Press, 1998 (*in Russian*), ISBN 5-89476-030-5

**202.** *Japan Statistical Yearbook*.- Tokyo: Ministry of Finance, 1993 (*in Japan and English*)

**203.** *Statistical Yearbook of Lithuania*.- Vilnius: Methodical Publishing Centre, 1998 (*in Lithuanian*)

**204.** *Statistical Yearbook of the Netherlands*.- The Hague: Statistics Netherlands Publications, 1995

**205.** *Statistical Abstract of the United States*.- Washington: The U.S. Government Printing Office, 1994

**206.** *Statistical Yearbook*.- New York: Publishing Division of UN, 1986 (*in English and French*)

**207. Muhacheva E., Rubinshtejn G.** *Mathematical Programming*.- Novosibirsk: Nauka Press, 1987 (*in Russian with extended English summary*)

**208.** *Forecasting of Social Processes in Socialist Society: Science as Object of Control*.- Kiev: Academy of Sciences of the Ukraine, Institute of Philosophy, 1969 (*in Russian with English summary*)

**209. Brusilovsky B.** *The Mathematical Models in Forecasting and Organization of Science*.- Kiev: Scientific thought, 1975 (*in Russian with extended English summary*)

**210.** *The fourth All-Union Symposium on Problems of Planning and Control by Scientific Probings and Designings*.- Moscow: Academy of Sciences of the USSR, 1977 (*in Russian with English summary*)

**211. Novikov E., Egorov V.** *Information and Researcher*.- Leningrad: Science Press, 1974 (*in Russian*)

**212. Gofman K.L. et al.** *Effectiveness of Science*.- Moscow: Science Press, 1984 (*in Russian*)

**213.** *Analysis of Regularities and Forecasting of Development of Science and Engineering*.- Kiev, 1967 (*in Russian with extended English summary*)

**214. Willems J.** From time series to linear system**/** *Automatica*, v. **22**, n. 5, 6 (1986); v. **23**, n. 1 (1987)

**215. Bek N.Z., Golenko D.P.** *Statistical Methods of Optimization in Economic Probings*.- Moscow: Statistics, 1971 (*in Russian with extended English summary*)

**216. Aladjev V.Z., Shishakov M.L.** Programming in package ***Maple V*** **// 2**nd Int. Conf. **"***Computer Algebra in Fundamental and Applied Reaserch and Education***".-** Minsk: Byelorussian Univ., 1999

**217. Aladjev V.Z., Shishakov M.** Automated working place of the mathematician **// 2**nd In. Conf. **"***Computer Algebra in Fundamental and Applied Reaserch and Education* **".-** Minsk: Byelorussian State University, 1999

**218. Kamps T.** *Diagram Design*.- Berlin: Springer-Verlag, 1999

**219. Hampel F. et al.** *Robust Statistics: The Approach Based on Influence Functions*.- N.Y.:Wiley, 1986

**220.** *Encyclopedia of Mathematics*, vol. 1.- London: Kluwer Academic Publisters, 1988, pp. 300-301.

**221. Aladjev V.Z.** *Interactive Encyclopedia of Cellular Automata*.- Tallinn: Tallinn Research Group & International Academy of Noosphere (*in preparation*)

**222. Aladjev V.Z. et al.** *Mathematical Theory of the Classical Homogeneous Structures*.- Gomel: BELGUT Press, 1998 (*in Russian with extended English summary*)
**223. Balakrishnan N., Chen W.** *CRC Handbook of Tables for Order Statistics*.- Berlin: Springer, 1997
**224. Dobrushin R. et al.** *Lecture on Probability Theory and Statistics*.- Paris: Springer, 1996
**225.** *CRC Standard Mathematical Tables and Formulae* / Ed. **D. Zwillinger.-** Berlin: Springer, 1995
**226.** *The Telecommunications Handbook* **/** Eds. **K. Terplan and P. Morreale.-** Berlin: Springer, 2000
**227.** *Computer Science. Newsletter*, no. **1-3.-** Berlin: Springer-Verlag, 2000
**228. Benker H.** *Practical Use of **MathCAD***.- Berlin: Springer-Verlag, 1999, ISBN 1-85233-166-6
**229. Mokhtari M.** ***MatLab** 5.2&5.3 and **Simulink** 2&3 for Engineers*.- Paris: Springer-Verlag, 2000
**230. Scott B.** ***Maple** for Environmental Sciences*.- Sidney: Springer-Verlag, 2000, ISBN 3-540-65826-2
**231.** *Statistical Models in Epidemiology, the Environment and Clinical Trials* **/** Eds. **M. Halloran** and **D. Berry.-** London: Springer, 2000, ISBN 0-387-98924-2
**232.** ***S-Plus:*** Academic Edition Version.- Cambridge: Mathsoft Inc., Springer, 2000
**233. Gander W., Hrebicek J.** *Solving Problems in Scientific Computing Using **Maple** and **MatLab***.- Zurich: Springer-Verlag, 1997, ISBN 3-540-61793-0
**234.** ***Macsyma 2.3.-*** Arlington: Macsyma Inc., Springer-Verlag, 1998
**235. Redfern D., Campbell C.** *The **MatLab 5** Handbook*.- Waterloo: Springer-Verlag, 1998
**236. Box G. et al.** *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*.- New York: John Wiley and Sons, 1978
**237. Devore J.** *Probability and Statistics for Engineering and the Sciences*.- N.Y.: Wadsworth Pub., 1995
**238. McCall R.** *Fundamental Statistics for the Behavioral Sciences*.- N.Y.: Harcourt Brace Jovanov, 1990
**239. Bettini C. et al.** *Temporal Databases with Multiple Granularities*.- Milan: Springer-Verlag, 2000
**240. Feuerlicht G.** *Object-Relational Database Management*.- Sydney: Springer-Verlag, 2000
**241. Kuper G., Libkin L., Paredaens J.** *Constraint Databases*.- Berlin: Springer-Verlag, 2000
**242.** ***Oracle WebDB. Version 2.0** Evaluation*.- Redwood Shores: Oracle Corp., CD, **#** C13319-01, 1999
**243. Bates D., Watts D.** *Non-linear Regression Analysis and its Applications*.- New York: Wiley, 1988
**244. Draper N., Smith H.** *Applied Regression Analysis*.- New York: John Wiley & Sons, 1981
**245. Ratkowsky D.** *Non-linear Regression Modeling*.- New York: Marcel Dekker, 1983
**246. Williams E.** *Regression Analysis*.- New York: John Wiley & Sons, 1959
**247. Box G., Jenkins G.** *Time Series Analysis, Forecasting and Practice*.- San Frisco: Holden Day, 1970
**248. Chatfield C.** *The Analysis of Time Series. Theory and Practice*.- London: Chapman and Hall, 1975
**249. Kendall M.** *Time Series*.- London: Griffin Press, 1973
**250. Nelson C.** *Applied Time Series for Managerial Forecasting*.- San Francisco: Holden Day, 1973
**251. Cleveland W.** *The Elements of Graphing Data*.- Wadsworth, California, 1985
**252. Box G., Draper N.** *Empirical Model Building and Response Surfaces*.- N.Y.: Wiley, 1987
**253. Fox J.** *Linear Statistical Models and Related Methods*.- N.Y.: John Wiley & Sons, 1984
**254. McCullagh P., Nelder J.** *Generalized Linear Models*.- London: Chapman and Hall, 1983
**255. Searle S.** *Linear Models*.- New York: John Wiley & Sons, 1971
**256. Cochran W.** *Sampling Techniques*.- N.Y.: John Wiley & Sons, 1977
**257. Gottman J.** *Time Series Analysis: A Comprehensive Introduction for Social Scientists*.- London: Cambridge University Press, 1981
**258. Conover W.** *Practical Nonparametric Statistics*.- New York: John Wiley & Sons, 1980
**259. Hollander M., Wolfe D.** *Nonparametric Statistical Methods*.- New York: John Wiley & Sons, 1973
**260. Neter J et al.** *Applied Statistics*.- Massachusetts: Allyn and Bacon, 1988

**261.** ***Maple 6.*** *The Standard for Analytical Computation*.- Waterloo: Waterloo Maple Inc., 2000
**262.** *The Seventh International Conference Computational Finance 2000*.- London: London Business School, June 2000
**263. Aladjev V.Z., Shishakov M., Trokhova T.** A workstation for solution of differential equations **// 3**rd Intern. Conf. "*Differential Equations and Applications*"**.-** Saint-Petersburg, 12-17 June 2000
**264. Keppel G.** *Design and Analysis: A Researcher's Handbook*.- N.J.: Prentice-Hall Inc., 1973
**265. Bradley J.** *Distribution-Free Statistical Tests*.- Englewood Cliffs, N.J.: Prentice-Hall Inc., 1968
**266. Lehmann E.** *Nonparametrics: Statistical Methods Based on Ranks*.- San Frisco: Holden-Day, 1975
**267. Everitt B.S.** *The Analysis of Contingency Tables*.- London: Chapman and Hall, 1977
**268. Zhukovskaja V., Mucnik I.** *The Factor Analysis in Social and Economical Investigations*.- Moscow: Higher School Press, 1976 (*in Russian with extended English summary*)
**269. Kurshakova Ju.S.** *Correlation and Regression Analysis in Practical Application*.- Novosibirsk, 1976 (*in Russian with extended English summary*)
**270. Kendall M., Buckland W.** *A Dictionary of Statistical Terms*.- London: Oliver & Boyd Press, 1960
**271. Mulhall M.G.** *The Dictionary of Statistics*.- Detroit: Book Tower Press, 1969
**272. Borovkov C.A.** *The English-Russian, Russian-English Dictionary on Probability Theory, Statistics and Combinatorics*.- Moscow-Philadelphia: SIAM, 1994
**273. Leser C.E.V.** *Econometric Techniques and Problems*.- London: Charles Griffin and Company, 1970
**274. Andersen P.K. et al.** *Statistical Models Based on Counting Processes*.- Copenhagen: Springer, 1997
**275. Jobson J.D.** *Applied Multivariate Data Analysis*. Vol. **1:** *Regression and Experimental Design*.- Berlin: Springer-Verlag, 1998
**276. Malliavin P.** *Stochastic Analysis*.- Paris: Springer-Verlag, 1997
**277. Moeschlin O. et al.** *Experimental Stochastics*.- Berlin: Springer-Verlag, 1998, CD + Booklet
**278. Simonoff J.S.** *Smoothing Methods in Statistics*.- N.Y.: Springer-Verlag, 1998, ISBN 0-387-94716-7
**279. Whittle P.** *Probability via Expectation*.- Cambridge: Springer-Verlag, 2000, ISBN 0-387-98955-2
**280. Capinski M., Zastawniak T.** *Probability via Problems*.- London-Berlin: Springer-Verlag, 2000
**281. Shorack G.S.** *Probability for Statisticians*.- Berlin: Springer-Verlag, 2000, ISBN 0-387-98953-6
**282.** ***S-Plus 2000:*** *Modern Statistics and Advanced Graphics*.- Cambridge: Mathsoft Inc., Springer, 2000
**283. Ivchenko G., Medvedev Ju.** *Mathematical statistics*.- Moscow: Higher School Press, 1984
**284. Seber G.S.** *Linear Regression Analysis*.- New York: John Wiley and Sons, 1977
**285. Kendall M., Stuart A.** *The Advanced Theory of Statistics*, v. 1-3.- London: Charles Griffin, 1970
**286. Aladjev V.Z., Shishakov M.L., Trokhova T.** Educational computer laboratory of the engineer **// 8**-th Byelorussian Mathemat. Conf., vol. 3.- Minsk, 2000, p. 154
**287. Aladjev V.Z. et al.** Simulation in software environment of the mathematical package ***Mapl*** **//** Intern. Conf. on Mathemat. Modelling.- Ukraine: Herson, 2000
**288. Aladjev V.Z. et al.** Workstation for solution of systems of differential equations **// 3**rd Intern. Conf. "*Differential Equations and Applications*"**.–** Saint-Petersburg, 2000
**289. Aladjev V.Z. et al.** Computer laboratory for engineering researches **//** Int. Conf. ***ACA-2000.–*** Saint-Petersburg, Russian Academy of Sciences, June 2000
**290. Aladjev V.Z. et al.** Workstation for mathematicians **//** Lithuanian Conf. "*TRANSPORT-2000*".- Vilnius, Vilnius Technical University, May 2000
**291. Paul W., Baschnagel J**. *Stochastic Processes: From Physics to Finance*.- Berlin: Springer, 1999
**292. Winkelmann R.** *Econometric Analysis of Count Data*.- Berlin: Springer-Verlag, 2000
**293. Sen A., Srivastava M.** *Regression Analysis: Theory, Methods and Applications*.- Chicago-Toronto: Springer-Verlag, 1997

**294. Härdle W. et al.** *Partially Linear Models*.- Berlin: Springer-Verlag, 2000, ISBN 3-7908-1300-1
**295. Back H.-H., Diday E.** *Analysis of Symbolic Data*.- Berlin-Paris: Springer-Verlag, 2000
**296. Isaac R.** *The Pleasures of Probability*.- Berlin: Springer-Verlag, 1996, ISBN 0-387-94415-X
**297.** *Data Analysis, Classification and Related Methods* **/** Eds. **H.A. Kiers et al.-** Berlin: Springer, 2000
**298.** *COMPSTAT 2000* – Proceedings in Computational Statistics **//** Eds. **J. Bethiehem et al.-** The Netherlands: Springer-Verlag, 2000
**299. Baltagi B.H.** *Econometrics*.- Berlin: Springer-Verlag, 1999
**300. Scmolck B.** *Omitted Variables Tests and Dynamic Specification*.- Zurich: Springer-Verlag, 2000
**301. Clegg F.** *Simple Statistics: A Course Book for the Social Sciences*.- London: Cambridge Univ., 1990
**302. Aladjev V.Z.** *Algebra Systems: A New Software Toolbox for* ***Maple.-*** Palo Alto: Fultus Publ., 2004
**303.** http**//**writers.fultus.com**/**aladjev**/**ebooks**/**victor_aladjev_new_software_toolbox_for_maple.pdf
**304.** http**//**www.macsyma.com**/**MXCompareMMA.html
**305. Hawley W.** *Foundations of Statistics*.- N.Y.: Saunders College Publishing, 1996
**306. Kirk R.** *Statistics: An Introduction*.- N.Y.: Saunders College Publishing, 1999
**307. Ross S.** *Introduction to Probability and Statistics for Engineers and Scientists*.- New York: Academic Press, 1999
**308. Roussas G.** *A Cource in Mathematical Statistics*.- New York: Academic Press, 1997
**309. Freund R., Wilson W.** *Regression Analysis*.- New York: Academic Press, 1998
**310. Hays W.** *Statistics*.- New York: Saunders College Publishing, 1993
**311. Newmark J.** *Statistics and Probability in Modern Life*.- N.Y.: Saunders College Publ., 1998
**312.** ***http://www.geocities.com/valadjev/Books/Books.htm***
**313.** *Encyclopedia of Mathematical Sciences*, vol. **1.-** Moscow: Soviet Enciclopedia, 1985, pp. 51 - 53
**314.** *Encyclopedia of Mathematics*.- The Hague: Kluwer Academic Publishing, 1995
**315. Cyganowski S., Kloeden P., Ombach J.** From Elementary Probability to Stochastic Differential Equations with Maple.- Berlin-London: Springer, 2002, 310 pp., ISBN 3-540-42666-3
**316.** CreditRisk in the Banking Industry **//** Eds. **M. Gundlach** and **F. Lehrbass.-** Berlin: Springer-Verlag, 2004, 369 p., ISBN 3-540-20738-4
**317. Elliot R., Kopp P.** *Mathematics of Financial Markets*.- Berlin: Springer-Verlag, 2004
**318. Shreve S.** *Stochastic Calculus for Finance: The Binomial Asset Pricing Model*.- Hamburg: Springer-Verlag, 2004, 187 p., ISBN 0-387-40100-8
**319. Shreve S.** *Stochastic Calculus for Finance: Continuous Time Models*.- Hamburg: Springer-Verlag, 2004, 550 p., ISBN 0-387-40101-8
**320. Meucci A.** *Risk and Asset Allocation*.- Greenwich: Springer-Verlag, 2004, ISBN 3-540-22213-8
**321. Musiela M., Rutkowski M.** *Martingale Methods in Financial Modelling*.- London: Springer-Verlag, 2004, 530 p., ISBN 3-540-20966-2
**322. Ruppert D.** *Statistics and Finance*.- London: Springer-Verlag, 2004, 473 p., ISBN 0-387-20270-6
**323. Franke J. et al.** *Statistics of Financial Markets*.- Paris: Springer-Verlag, 2004, ISBN 0-387-20270-6
**324. Roman S.** *Introduction to the Mathematics of Finance: From Risk Management to Options Pricing*.- London: Springer-Verlag, 2004, 360 p., ISBN 0-387-21375-9
**325.** *Statistical Tools in Finance and Insurance* **// P. Cizet at al.-** London: Springer-Verlag, 2004
**326. Skorokhod A.** *Basic Principles and Applications of Probability Theory*.- Berlin: Springer, 2004
**327. Tsirelson B., Werner W.** *Lectures on Probability Theory and Statistics*.- Paris: Springer-Verlag, 2004, 200 p., ISBN 3-540-21316-3
**328. Tavare S.** *Lectures on Probability Theory and Statistics*.- Paris: Springer-Verlag, 2004, 315 p.

**329. Catoni O.** *Statistical Learning Theory and Stochastic Optimization*.- Paris: Springer-Verlag, 2004
**330. Anderson C., Loynes R.** *The Teaching of Practical Statistics*.- N.Y.: John Wiley & Sons, 1987
**331. Andrews D. , Herzberg A.** *Data, A Collection of Problems from Many Fields for the Student and Research Workers*.- New York: Springer-Verlag, 1985
**332. Barnett V.** *Advanced Level Studies: Statistics, Statistical Background*.- Sheffield: Center for Statistical Education, University of Sheffield, 1985
**333. Bibby J.** *History of Teaching Statistics*.- London: John Bibby, 1986
**334.** *Guidelines for the Teaching of Statistics* **//** Ed. **G. Burrill.-** Alexandria, Virginia: ASA, 1991
**335.** *Teaching and Using Statistics* **//** Ed. **N. Davies.-** London: Royal Statistical Society, 1993
**336. Gal I., Garfield J.** *The Assessment Challenge in Statistics*.- Voorburg, The Netherlands: ISI, 1997
**337. Graham A.** *Teach Yourself Statistics*.- London: Hodder and Soughton, 1994
**338.** *Teaching Statistics at Its Best //* Ed. **D. Green.-** Sheffield: Teaching Statistics Trust, 1993
**339.** *A Handbook of Small Data Sets* **//** Eds. **D. Hand et al.-** London. Chapman and Hall, 1994
**340. Hawkins A. et al.** *Teaching Statistical Concepts*.- London: Longman, 1992
**341. Noggle J.** *Practical Curve Fitting and Data Analysis, Software and Self-instruction for Scientists and Engineers*.- London: Horwood, 1993
**342.** *Teaching of Statistics in the Computer Age* **//** Eds. **L. Räde** & **T. Speed.-** Kent: Chartwell, 1985
**343. Abelson R.** *Statistics as Principled Argument*.- NJ: Lawrence Erlbaum Associates, 1995
**344. Argyrous G.** *Statistics for Social and Health Research*.- London: Sage Publications, 2000
**345. Brown F. et al.** *Statistical Concepts: A Basic Program*.- NY: HarperCollins College Publ., 1995
**346. Craft J.** *Statistics and Data Analysis for Social Workers*.- Itasca, IL: F.E. Peacock Publ., 1990
**347. Aladjev V.Z., Haritonov V.N.** *General Theory of Statistics.* – Palo Alto: Fultus Publishing, 2004
**348.** *Statistical Tools for Finance and Insurance* **//** Eds. **P. Cizek, W. Hardle, R. Weron.-** Berlin: Springer, 2004, ISBN 3-540-22189-1
**349. Good R.** *Permutation, Parametric, and Bootstrap Tests of Hypotheses*.- N.Y.: Springer, 2005
**350. Scherer B, Martin R.** *Introduction to Modern Portfolio Optimization with* **NUOPT** *and* ***S-Plus.-*** Berlin: Springer, 2005, ISBN 0-387-21016-4
**351.** *New Developments in Classification and Data Analysis* **//** Eds. **M. Vichi et al.-** Roma: Springer, 2005, ISBN 3-540-23809-3
**352. Zivot, Wang J.** *Modeling Financial Time Series with* **S-Plus.-** N.Y.: Springer, 2004
**353.** *Statistical Tools for Finance and Insurance* **/** *Eds.* **P. Cizek et al.-** Berlin: Springer, 2004
**354. Dekking F.M. et al.** *A Modern Introduction to Probability and Statistics*.- Berlin: Springer, 2005.
**355. Никитин А.В., Гачко Г.А., Слободянюк А.И.** ***VISUAL BASIC:*** Учебный курс.- Гродно: Гродненский Государственный университет, 2005, 364 с., ISBN 985-417-643-6
**356. Ефимова М.Р.** *Общая теория статистики*. Учебник.- М.: Изд-во: ИНФРА-М, 2005
**357. Елисеева И.И., Юзбашев М.М.** *Общая теория статистики*. Учебник.- М.: Изд-во: "Финансы и статистика", 2005, ISBN 5-279-01956-9
**358. Кожухарь Л. И.** *Основы общей теории статистики*.- М.: Изд-во: "Финансы и статистика", 1999, ISBN 5-279-02017-6
**359. Бендина Н. В.** *Общая теория статистики*. Пособие для подготовки к экзаменам.- М.: Изд-во: Приор, 1999, ISBN 5-7990-0272-5
**360. Ефимова М.Р., Ганченко О.И., Петрова Е.В.** *Практикум по общей теории статистики*.- М.: Изд-во: "Финансы и статистика", 2006, ISBN 5-279-02555-0

# Список Рисунков

# Список Таблиц

# Профессиональные статистические и математические организации

Agency for Statistics of Bosnia and Herzegovina; spopovic@bih.net.ba
Agency on Statistics of the Republic of Kazakhstan; kazstat@mail.banknet.kz
Albania: Institute of Statistics – http://www.instat.gov.al
American Mathematical Society (**AMS**); ams@math.ams.org
American Society for Quality (**ASQ**) – http://www.asq.org
American Statistical Association (**ASA**) – http://www.amstat.org
Applied Mathematics Society, Department of Mathematics and Statistics; gac@cs.sfu.ca
Argentina: National Institute of Statistics and Censuses – http://www.indec.mecon.ar
Armenia: Ministry of Statistics – http://www.armstat.am
Association for Computing Machinery (**ACM**) – http://info.acm.org
Australia: Australian Bureau of Statistics – http://www.abs.gov.au
Austria: Central Statistical Office – http://www.statistik.at
Azerbaijan: State Statistical Committee – http://www.azeri.com/goscomstat
Belgium: National Institute of Statistics – http://www.statbel.fgov.be
Bernoulli society for mathematical statistics and probability – www.cbs.nl/isi/BS/bshome.htm
Brazil: Brazilian Institute of Geography and Statistics (**IBGE**) – http://www.ibge.gov.br
Bulgaria: National Statistical Institute – http://www.nsi.bg
Canada: Statistics Canada – http://www.statcan.ca
Canadian Mathematical Society; exsmc@acadvm1.uottawa.ca
Central Bureau of Statistics of Croatia – http://rusan@dzs.hr
Central Bureau of Statistics of Israel; yahav@cbs.gov.il
Central Office of Statistics, Malta; cos@magnet.mt
Central Statistical Bureau of Latvia; Azigure@csb.lv
Central Statistical Office of Poland; t.toczynski@stat.gov.pl
Central Statistics Office of Ireland; dg@cso.ie
Chile: National Institute of Statistics (**INE**) – http://www.ine.cl
China: National Bureau of Statistics – http://www.stats.gov.cn/english/index.htm
Classification Society of North America (**CSNA**) – http://www.pitt.edu/~csna
Consortium for Mathematics and Its Applications (**COMAP**); info@comap.com
Czech Republic: Czech Statistical Office – http://www.czso.cz/eng/angl.htm
Denmark: Statistics Denmark– http://www2.dst.dk/internet/startuk.htm
Departament d'Estudis I d'Estadistica, Ministeri de Finances; servest@andorra.ad

Department of Statistics, Liechtenstein; Christian.Brunhart@avw.llv.li
ECE Statistical Division; sari.saleh@unece.org
Econometric Society – http://www.econometricsociety.org
Estonia: Statistical Office of Estonia – http://www.stat.ee
European Community: EUROSTAT – http://europa.eu.int/comm/eurostat
European Network for Business and Industrial Statistics (**ENBIS**) - http://www.enbis.org
Federal Statistical Office, Germany; johann.hahlen@statistik–bund.de
Federal Statistical Office, Yugolsavia; zivkovic@szs.gov.yu
Finland: Statistics Finland – http://www.stat.fi/index_en.html
Finnish Statistical Society – http://www.stat.fi/sts/index_en.html
Food and Agricultural Organization (**FAO**) – http://www.fao.org/
German region of the international biometric society –www.dkfz–heidelberg.de/biostatistics/ibs
Germany: Federal Statistical Office – http://www.destatis.de/e_home.htm
Greece: National Statistical Service – http://www.statistics.gr
Hungarian Central Statistical Office; tamas.mellar@ksh.x400gw.itb.hu
Hungary: Hungarian Central Statistical Office – www.ksh.hu/pls/ksh/docs/index_eng.html
India: Department of Statistics (General Statistics) – http://mospi.nic.in
Indonesia: Central Bureau of Statistics – http://www.bps.go.id
Industrial Mathematics Society; P.O. Box 159, Roseville, MI 48066
Institut de la Statistique; mekonomi@instat.gov.al
Institut National de Statistique, Italie; PRES@ISTAT.IT
Institut National de Statistique; ahadjiiski@nsi.bg
Institut National de Statistique; claude.cheruy@statbel.mineco.fgov.be
Institute for Scientific Information (**ISI**), Philadelphia, PA 19104, U.S.A.
Institute of Mathematical Statistics (**IMS**) – http://www.imstat.org
Instituto Nacional de Estadística, Spain; pmguzman@ine.es
International Association for Official Statistics (**IAOS**) - http://www.stat.fi/iaos
International Association for Statistical Computing (**IASC**) – http://www.cbs.nl/isi/iasc.htm
International Association for Statistical Education (**IASE**) – www.stat.auckland.ac.nz/~iase/
International Association of Survey Statisticians (**IASS**) – http://www.cbs.nl/isi/iass/index.htm
International Biometric Society (**IBS**) – http://www.tibs.org
International Chinese Statistical Association (**ICSA**) – http://www.icsa.org
International Labour Organization (**ILO**) – http://laborsta.ilo.org
International Mathematics Union (**IMPA**); imu@impa.br
International Monetary Fund (**IMF**) – http://www.imf.org
International Society for Bayesian Analysis (**ISBA**) – http://www.bayesian.org
International Statistical Institute (**ISI**) – http://www.cbs.nl
Ireland: Central Statistics Office – http://www.cso.ie
Irving Fisher Society for Financial and Monetary Statistics – http://www.cbs.nl/isi/fisher.htm
Israel: Central Bureau of Statistics – http://www.cbs.gov.il/engindex.htm

Italy: National Statistical Institute – http://www.istat.it
Japan: Japanese Statistics Bureau – http://www.stat.go.jp
Joint Policy Board for Mathematics; jpbm@math.umd.edu
Latvia: Central Statistical Bureau of Latvia – http://www.csb.lv/avidus.cfm
l'INSEE, France; paul.champsaur@insee.fr
Lithuania: Statistics Lithuania – http://www.std.lt/web/main.php
Luxembourg: National Institute of Statistics and Economic Studies – http://statec.gouvernement.lu
Mathematical Association of America (**MAA**); maahq@maa.org
Mexico: National Institute of Statistics, Geography, and Informatics – http://www.inegi.gob.mx
Ministry of Statistics and Analysis, Byelorussia; svet@domhos.belpak.minsk.by
Moldova: Department for Statistics and Sociology – http://www.statistica.md/?lang=en
National Association of Mathematicians; nam@ecsvax.uncecs.edu
National Commission for Statistics of Romania; romstat@cns.ro
National Institute of Statistics and Economic Studies – http://www.insee.fr/en/home/home_page.asp
National Statistical Committee of the Kyrgyz Republic; 311@nsc.bishkek.su
National Statistical Institute, Portugal; INE@mail.telepac.PT
National Statistical Service of Greece; general.secretary@statistics.gr
National Statistical Service of the Republic of Armenia; armstat@sci.am
National Statistician, Statistics Denmark; jpl@dst.dk
Netherlands: Central Bureau of Statistics (**CBS**) – http://www.cbs.nl/en
New Zealand: Statistics New Zealand – http://www.stats.govt.nz
Norway: Statistics Norway – http://www.ssb.no/www–open/english
Office for National Statistics, United Kingdom; len.cook@ons.gov.uk
Office of Management and Budget, Executive Office of the President; kwallman@omb.eop.gov
Organization for Economic Cooperation and Development (**OECD**) – http://www.oecd.org
Organization of Economic Cooperation and Development (**OECD**) – http://www.oecd.org/statistics
Poland: Central Statistical Office – http://www.stat.gov.pl/english/index.htm
Royal Statistical Society (**RSS**) – http://www.rss.org.uk
Royal Statistical Society, London, Great Britain; rss@rss.org.uk
Russia: Russian State Committee for Statistics – http://www.gks.ru/eng
Service Central de la Statistique et des Etudes Economiques (**STATEC**); robert.weides@statec.etat.lu
Singapore: Department of Statistics – http://www.singstat.gov.sg
Slovakia: Statistical Office – http://www.statistics.sk/webdata/english/index2_a.htm
Society for Computational Economics – http://wueconb.wustl.edu/sce/
Society of Industrial and Applied Mathematics (**SIAM**); siam@siam.org
Spain: National Institute of Statistics – http://www.ine.es
State Committee of the Russian Federation on Statistics; sokolin@gks.ru
State Department for Statistics of Georgia; soceinf@iberiapac.ge

State Institute of Statistics, Turkey; sefik.yildizeli@die.gov.tr
State Statistical Agency at the Government of the Republic of Tajikistan; tad@stat.td.silk.org
State Statistical Committee of Azerbaijan; azstat@azeri.com
State Statistics Committee of Ukraine; minstat@minstat.kiev.ua
Statistical Committee of **CIS** – http://www.cisstat.com/
Statistical Division Economic Commission for Europe; paolo.garonna@unece.org
Statistical Office of Estonia; rein.veetousme@stat.ee
Statistical Office of Macedonia; svetlana@stat.gov.mk
Statistical Office of the European Communities (**Eurostat**) – http://europa.eu.int/comm/eurostat
Statistical Office of the European Communities; yves.franchet@eurostat.cec.be
Statistical Office of the Republic of Slovenia; tomaz.banovec@gov.si
Statistical Office of the Slovak Republic; mach@statistics.sk
Statistical Office of the United Nations; statistics@un.org
Statistical Service of Cyprus; cydsr@cytanet.com.cy
Statistical Society of Australia Inc. (**SSAI**) – http://www.statsoc.org.au
Statistical Society of Canada – ***http://www.ssc.ca***
Statistical Society of Canada; roger@uvvm.uvic.ca
Statistics Austria; ewald.kutzenberger@oestat.gv.at
Statistics Canada; fellegi@statcan.ca
Statistics Department International Monetary Fund; CCARSON@IMF.ORG
Statistics Directorate **OECD,** France; KINCANNON@OECD.ORG
Statistics Division United Nations; HABERMANN@UN.ORG
Statistics Finland; timo.relander@stat.fi
Statistics Iceland; hallgrimur.snorrason@statice.is
Statistics Lithuania; statistika@mail.std.lt
Statistics Netherlands; rnrt@CBS.NL
Statistics Norway, slo@ssb.no
Statistics Sweden; svante.oberg@scb.se
Sweden: Statistics Sweden – http://www.scb.se
Swiss Federal Statistical Office; carlo.malaguerra@bfs.admin.ch
Switzerland: Swiss Federal Statistical Office – http://www.statistik.admin.ch/eindex.htm
The Department for Statistical and Sociological Research, Republic of Moldova; dass@moldova.md
Ufficio programmazione economica e Centro Elaborazione dati e statistica; progecon@omniway.sm
Ukraine: State Committee of Statistics – http://www.ukrstat.gov.ua
UN Economic Commission for Europe (**UN/ECE**) – http://www.unece.org
UN Statistics Division – http://www.un.org/Depts/unsd
United Arab Emirates: Ministry of Planning – http://www.uae.gov.ae/mop/
United Kingdom: Office of National Statistics – http://www.statistics.gov.uk
United Nations Industrial Development Organization (**UNIDO**) – www.unido.org/doc/3474
United Nations Statistics Division – http://unstats.un.org/unsd/

United Nations, Division for Sustainable Development, dsd@un.org
United States Statistical Agencies – http://www.fedstats.gov
Uruguay: Statistical Department – http://www.ine.gub.uy
Uzbekistan: Ministry of Macroeconomy and Statistics – http://www.gov.uz
World Bank – http://www.worldbank.org
World Health Organization (**WHO**) – http://www3.who.int/whosis/menu.cfm
World Trade Organization(**WTO**) – http://www.wto.org

# Международные периодические издания по статистике

Allgemeines Statistisches Archiv – Journal of the German Statistical Society, ISSN 0002–6018 (*print*)
American Behavioral Scientist – ISSN 0002–7642
American Journal of Political Science – ISSN 0092–5853
American Journal of Sociology – ISSN 0002–9602
American Political Science Review – ISSN 0003–0554
American Sociological Review – ISSN 0003–1224
Annual Review of Sociology – ISSN 0360–0572
Applied Measurement in Education – ISSN 0895–7347
Behavior Research Methods, Instruments, & Computers – ISSN 0743–3808
Behaviormetrika – ISSN 0385–7417
British Journal of Mathematical and Statistical Psychology – ISSN 0007–1102
British Journal of Political Science – ISSN 0007–1234
British Journal of Social Psychology – ISSN 0144–6665
Business Statistics – Washington, U.S. Department of Commerce, Bureau of Economic Analysis
Business Statistics of the United States – Bernan Press. Annual
Calcutta Statistical Association Bulletin, Department of Statistics, Calcutta University (India)
Computational Statistics – ISSN 0943–4062 (*print*)
CPI Detailed Report – Washington, U.S. Department of Labor, Bureau of Labor Statistics
Cross–Cultural Research; The Journal of Comparative Social Science – ISSN 1069–3971
Current Contents; Social and Behavioral Sciences – ISSN 0092–6361
Current Index to Statistics, Applications, Methods and Theory – Washington, Annual
Decision Sciences, Stanford University (USA)
Economic Bulletin – German Institute for Economic Research, ISSN 0343–754X (*print*)
Economic Survey of Europe – Geneva, Economic Commission for Europe
Economic Systems – Osteuropa–Institut München in collaboration with the European Association for Comparative Economic Studies (**EACES**), ISSN 0939–3625 (*print*)
Economic Theory – Official Journal of the Society fot the Advancement of Economic Theory, ISSN 0938–2259 (*print*), ISSN 1432–0479 (*electronic*)
Economics of Governance – ISSN 1435–6104 (*print*), ISSN 1435–8131 (*electronic*)
Educational and Psychological Measurement – ISSN 0013–1644
Empirical Economics – ISSN 0377–7332 (*print*), ISSN 1435–8921 (*electronic*)
Evaluation and Program Planning; International Journal – ISSN 0149–7189

Evaluation; The International Journal of Theory – ISSN 1356–3890
Field Methods –ISSN 1525–822X
Finance and Stochastics – Bonn, Germany, ISSN 0949–2984 (*print*), ISSN 1432–1122 (*electronic*)
Group Dynamics – ISSN 1089–2699
IEEE Transactions on Reliability, Institute of Electrical and Electronics Engineers (New York)
International Financial Statistics – Washington, Statistics Bureau, International Monetary Fund
International Journal of Forecasting – ISSN 0169–2070
International Journal of Market Research – ISSN 0025–3618
International Journal of Qualitative Methods – http://www.ualberta.ca/~ijqm/
International Journal of Social Research Methodology – ISSN 1364–5579
International Social Science Journal – ISSN 0020–8701
Journal of Applied Measurement – ISSN 1529–7713
Journal of Applied Statistical Science, Nova Science (Commack, NY)
Journal of Business & Economic Statistics – ISSN 0735–0015
Journal of Classification – ISSN 0176–4268
Journal of Consumer Research – ISSN 0093–5301
Journal of Educational and Behavioral Statistics – ISSN 1076–9986
Journal of Educational Measurement – ISSN 0022–0655
Journal of Evolutionary Economics – ISSN 0936–9937 (*print*), ISSN 1432–1286 (*electronic*)
Journal of Marketing Research – ISSN 0022–2437
Journal of Marketing; American Marketing Association – ISSN 0022–2429
Journal of Official Statistics – ISSN 0282–423X
Journal of Population Economics – ISSN 0933–1433 (*print*), ISSN 1432–1475 (*electronic*)
Journal of the Academy of Marketing Science – ISSN 0092–0703
Journal of the American Statistical Association – ISSN 0162–1459
Journal of the European Mathematical Society – the European Mathematical Society, ISSN 1435–9855 (*print*), ISSN 1435–9863 (*electronic*); e–mail: jems@mis.mpg.de
Journal of the International Actuarial Association (Belgium)
Journal of the Japan Statistical Society, Institute of Statistical Mathematics (Tokyo)
Journal of the Royal Statistical Society –ISSN 0964–1998, ISSN 0039–0526
Main Economic Indicators – Paris, Organization for Economic Co–operation and Development
Marketing Theory – ISSN 1470–5931
Mathematical Methods of Statistics, Allerton Press (New York)
Mathematical Population Studies – ISSN 0889–8480
Methoden und Instrumente der Sozial–wissenschaften – ISSN 0176–4446
Metrika – International Journal for Thepretical and Applied Statistics, ISSN 0026–1335 (*print*), ISSN 1435–926X (*electronic*)
Monthly Bulletin of Statistics – NY: Statistical Office of the United Nations
Multivariate Behavioral Research – ISSN 0027–3171

**NB:** See electronic versions of Springer journals in web–site WWW: http://link.springer–ny.com

Netherlands Official Statistics (CBS) – ISSN 0920–2048
Organizational Research Methods – ISSN 1094–4281
Papers in Regional Science – The Journal of the Regional Science Association International, ISSN 1056–8190 (*print*), ISSN 1435–5957 (*electronic*)
Psychometrika – ISSN 0033–3123
Qualitative Health Research – ISSN 1049–7323
Qualitative Inquiry – ISSN 1077–8004
Qualitative Market Research – ISSN 1352–2752
Qualitative Sociology – ISSN 0162–0436
Quality & Quantity – ISSN 0033–5177
Review of Economic Design – ISSN 1434–4742 (*print*), ISSN 1434–4750 (*electronic*)
Revue de Statistique Appliquee - **CERESTA,** Inst de Stat des Université de Paris (France)
Scandinavian Journal of Statistics; Theory and Applications – ISSN 0303–6898
Selecta Statistica Canadiana, Department of Mathematics, McMaster University (Canada)
Social Choice and Welfare – ISSN 0176–1714 (*print*), ISSN 1432–217X (*electronic*)
Social Indicators Research – ISSN 0303–8300
Social Networks; International Journal of Structure Analysis – ISSN 0378–8733
Social Science Computer Review – ISSN 0894–4393
Social Science Research – ISSN 0049–089X
Sociological Methods and Research – ISSN 0049–1241
Staff Papers International Monetary Fund – ISSN 0020-7635
Standartizacia, Sertifikazia, Metrologia – ISSN 1310-0831
Standarty i kacestvo – ISSN 7087 8
Statistica – ISSN 0039-0380
Statistica – ISSN 0390-590X
Statistica Neerlandica – ISSN 0039-0402
Statistical Abstracts of United States – ISSN 0081-4741
Statistical Journal of the United Nations for Europe – ISSN 0167-8000
Statistical Methods in Medical Research – ISSN 0962-2802
Statistical News – ISSN 0204-563X
Statistical Papers – ISSN 0932–5026 (*print*), ISSN 1435–151X (*electronic*)
Statistical Science – ISSN 0883-4237
Statistical Theory and Methods Abstracts on CD ROM – ISSN 0039-0518
Statistics - A Journal of Theoretical and Applied Statistics – ISSN 0233-1888
Statistics and Computing – ISSN 0960-3174
Statistics and Decisions – ISSN 0721-2631
Statistics and Probability Letters – ISSN 0167-7152
Statistics in Medicine – ISSN 0277-6715
Statistische Nachrichten – ISSN 0029-9960
Structural Equation Modeling; A Multidisciplinary Journal – ISSN 1070–5511

Survey Methodology; A Journal published by Statistics Canada – ISSN 0714–0045
Teaching Statistics, Department of Probability and Statistics, University of Sheffield (UK)
The American Statistician; American Statistical Association – ISSN 0003–1305
The British Journal of Sociology – ISSN 0007–1315
The Journal of Applied Behavioral Science – ISSN 0021–8863
The Journal of Mathematical Sociology – ISSN 0022–250X
The Qualitative Report – http://www.nova.edu/ssss/QR/
The Review of Economics and Statistics – Cambridge, Dept. of Economics, Harvard University

# Список основных используемых обозначений

**AAI** – средне арифметический индекс
**ACC** – коэффициент автокорреляции
**AGI** – средне геометрический индекс
**AHI** – средне гармонический индекс
**Aidx** – агрегатный индекс
**AUI** – средне взвешенный индекс
**BI** – базисный индекс
**CC** – коэффициент корреляции
**CI** – цепной индекс
**CPI** – индекс потребительских цен
**CR** – корреляционное отношение
**DF** – определяющая функция
**ErCC** – средне квадратичная ошибка
**MCC** – множественный коэффициент корреляции
**Me** – медиана
**MFRM** – многофакторная модель регрессии
**MI** – модальный интервал
**Mo** – мода
**MSA** – многофакторный статистический анализ
**PCC** – частный коэффициент корреляции
**АРМ** – автоматизированное рабочее место
**БД** – база данных
**ВВП** – величина выполнения плана
**ВД** – величина динамики
**ВИ** – величина интенсивности
**ВН** – выборочное наблюдение
**ВПЗ** – величина планового задания
**ВР** – вариационный ряд
**ВТ** – вычислительная техника
**ДС** – департамент статистики
**ЗБЧ** – закон больших чисел
**КЕС** – Конференция Европейских Статистиков
**ЛМР** – линейная модель регрессии
**МВ** – малая выборка
**МНК**– метод наименьших квадратов
**МО** – математическое ожидание
**МСА** – многомерный статистический анализ
**МТОС**– математическая теория однородных структур
**НД** – национальный доход
**НМР**– нелинейная модель регрессии
**ПК** – персональный компьютер
**ППП** – пакет прикладных программ
**ПС** – программное средство
**СКО** – средне квадратичное отклонение
**СНР** – стандартное нормальное распределение
**СУБД** – система управления базами данных
**ТСИ** – теория статистических игр
**ТТГ** – Таллиннская Творческая Группа
**ЦПТ** – центральная предельная теорема
**ЦСУ** – Центральное Статистическое Управление

# Index