

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Государственное образовательное учреждение  
высшего профессионального образования

«ЛИПЕЦКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»

А.В. Галкин, С.А. Жбанов, Е.В. Кузнецова

# **КОМПЬЮТЕРНЫЙ ПРАКТИКУМ ПО ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ**

**Учебное пособие**



Липецк  
Издательство ЛГТУ  
2011

УДК 519.6  
Г161

Рецензенты:

кафедра прикладной математики и информационных технологий  
Липецкого государственного педагогического университета,  
Седых И. А., кандидат физ.-мат. наук,  
заведующая кафедрой математических, естественнонаучных и  
экономических дисциплин Липецкого института права и экономики

**Галкин, А.В.**

Г161 Компьютерный практикум по теории вероятностей и математической статистике [Текст]: учеб. пособие / А.В. Галкин, С.А.Жбанов, Е.В. Кузнецова. – Липецк: Изд-во ЛГТУ, 2011. – 80 с.

**ISBN**

Настоящее учебное пособие содержит необходимые сведения и формулы теории вероятностей и математической статистики, изложение основ технологии анализа данных на компьютере средствами Excel и системы STATISTICA, задания для самостоятельного изучения, решения типовых задач, материал для самостоятельного изучения, задания для самостоятельной работы и лабораторных работ.

Материал структурирован таким образом, чтобы обеспечить возможность построения индивидуальной траектории обучения в зависимости от уровня подготовленности обучаемых и педагогических задач, поставленных преподавателем.

Пособие предназначено для организации учебного процесса и самостоятельной работы студентов технического университета очной и очно-заочной форм обучения при изучении курса теории вероятностей и математической статистики.

**ISBN**

© Галкин А.В., Жбанов С.А., Кузнецова Е.В., 2011

© Липецкий государственный  
технический университет, 2011

Учебное издание

**Компьютерный практикум по теории вероятностей и  
математической статистике**

**Учебное пособие**

**Галкин Александр Васильевич**

**Жбанов Сергей Александрович**

**Кузнецова Елена Васильевна**

Редактор М.Ю. Копытина

Подписано в печать

2011. Формат 60х84 1/16. Бумага офсетная.

Ризография.

Объем 5,0 печ. л. Тираж 200 экз. Заказ № .

Издательство Липецкого государственного технического университета.

Полиграфическое подразделение Издательства ЛГТУ.

398600 Липецк, ул. Московская, 30.

## **Введение**

В условиях развития информационного общества в профессиональной деятельности специалиста можно выделить два взаимосвязанных направления: применение современного программного обеспечения, а также применение и исследование математических методов и моделей объектов, систем, процессов и технологий, ибо компьютеризация и математизация являются важной приметой современной жизни.

Учитывая этот факт, а также присутствие случайностей в природных процессах, технике, экономике, социальной сфере и других отраслях человеческой деятельности, можно сделать вывод о том, что развитие производства и информационных технологий поставило перед человечеством ряд задач и проблем, решить которые под силу только специалисту, имеющему основательную стохастическую подготовку. Действительно, вероятностные идеи играют важную роль в современном научном познании, формировании научной картины мира, наиболее адекватно отражающей изменчивость, нестабильность, риски информационного общества. Благодаря развитию стохастических методов и моделей применение математики расширило свои границы от изучения простых и точных зависимостей до моделирования сложных явлений в экономике, социологии, психологии, медицине, образовании и других сферах, где действует человек, обладающий свободой воли и свободой выбора.

В связи с этим в преподавании курса теории вероятностей и математической статистики важным компонентом является компьютерный практикум. Не случайно стандарт нового поколения для ряда направлений предусматривает обязательное включение лабораторного практикума при изучении вероятностных разделов математики.

Цели компьютерного практикума по теории вероятностей и математической статистике:

- содействие пониманию вероятностной природы изучаемых объектов, более глубокое проникновение в сущность случайных явлений;
- активное осмысленное усвоение теоретических положений, вероятностных понятий и законов;
- формирование основных умений, необходимых для анализа и обработки данных с применением компьютера, освоение особенностей статистического вывода;

- приобретение навыков стохастического моделирования.

Предполагается, что наиболее эффективным является сочетание таких организационных форм, как работа с преподавателем, самостоятельная работа (с консультациями преподавателя), домашняя работа и лабораторная работа, представляющая самостоятельное исследование с последующей защитой. Лабораторные работы предусматривают как использование известных программных продуктов (Excel, Statistica и др.), так и создание студентами собственных программ с применением стохастических функций.

Необходимо отметить два принципа использования ЭВМ в учебном процессе: как средства вычисления (для расчетов значений вероятности при решении задач или для вычисления числовых характеристик выборки при анализе данных), а также как инструмента познания (в процессе стохастического компьютерного моделирования). В первом случае удастся избежать рутинных вычислений. Во втором – открывается перспектива как в познавательном плане, так и для осознания связи информатики с математикой, естественными и гуманитарными науками, что способствует развитию интуиции и исследовательских навыков в ситуациях неопределенности и выбора, активизирует познавательную деятельность.

Компьютерный практикум в курсе стохастики позволяет реализовать дидактические принципы:

- наглядность (например, использование преимуществ графического анализа);
- доступность, посильность и индивидуализация обучения (возможность построения занятия и формирование заданий с учетом уровня подготовки студентов и образовательных задач, поставленных преподавателем);
- сознательность и активность (задания для самостоятельной работы, ссылки на источники информации, исследовательские задачи).

Как известно, современный вузовский учебник должен представлять собой модель учебного процесса, то есть, оставаясь средством познания, учебная книга все больше принимает на себя роль организатора и руководителя процесса обучения. В связи с этим материал в данном учебном пособии структурирован так, чтобы изучению темы было посвящено отдельное занятие. Это особенно важно для системы заочного обучения, поскольку у студентов возникают проблемы с организацией самостоятельной работы.

## Занятие 1

### Основы обработки и анализа данных

В настоящее время невозможно представить принятие решений без использования методов математического моделирования и анализа данных. Теоретические основы рассматривались многими авторами [4, 8, 11].

#### Цели занятия

**Знать** – определения понятий «генеральная совокупность» и «выборка», сущность выборочного метода, определения и формулы вычисления основных числовых характеристик выборки.

**Уметь** – производить группировку и сортировку данных, вычислять и анализировать основные числовые характеристики выборки.

**Владеть** – методами первичной обработки и графического анализа данных средствами *Excel* и *STATISTICA*.

Вся подлежащая исследованию совокупность объектов называется *генеральной совокупностью*. Та часть объектов из генеральной совокупности, которая попала на проверку (исследование), называется *выборочной совокупностью* (*выборкой*). Число объектов в совокупности называется ее *объемом*.

Обработка данных начинается с их упорядочивания по возрастанию и **группировки**. Возможно представление выборки в несгруппированном виде т.е. в виде вариационного ряда, когда все значения признака располагаются в порядке возрастания (в этом случае значения называются *вариантами*), или в сгруппированном виде (дискретная или интервальная выборка для непрерывного распределения).

Перечень вариант  $x_i$  и соответствующих им частот  $n_i$  (относительных частот  $\omega_i = n_i/n$ ) называется *статистическим распределением выборки* или *статистическим рядом*. Сумма всех частот равна объему выборки, а сумма относительных частот равна 1.

Статистическое распределение выборки можно задать также в виде последовательности интервалов и соответствующих им частот (количество вариант, попавших в этот интервал).

**Статистической оценкой** неизвестного параметра теоретического распределения называют функцию от наблюдаемых значений случайной величины.

**Точечной** называют оценку, которая определяется одним числом. Точечными оценками являются выборочная средняя  $\bar{x}$  (характеристика положения); выборочная дисперсия  $D_B$ , выборочное среднее квадратическое отклонение (СКО)  $\hat{\sigma} = \sqrt{D_B}$  (характеристики рассеяния). К точечным оценкам относятся также выборочные мода, медиана, асимметрия и эксцесс.

**Замечание 1.** Вместо термина «выборочное СКО» в ряде источников применяется термин «стандартное отклонение».

**Замечание 2.** Выборочная средняя  $\bar{x}$ , вычисленная по эмпирическим данным, является случайной величиной. Оценка СКО случайной величины  $\bar{x}$  называется стандартной ошибкой.

**Характеристики положения** определяют положение центра эмпирического распределения.

**1) Выборочная средняя** представляет собой среднее арифметическое значение признака выборочной совокупности:  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  – для несгруппированных данных;  $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i n_i$  – для сгруппированных данных, где  $\sum_{i=1}^k n_i = n$  – объем выборки, а  $k$  – количество интервалов или значений признака для сгруппированных данных.

**2) Выборочная мода  $\bar{M}_0$ :** модой дискретного вариационного ряда является варианта, имеющая наибольшую частоту; мода интервального вариационного ряда определяется по формуле  $\bar{M}_0 = x_0 + h \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}$  (здесь  $i$  – номер;  $x_0$  – начало;  $h$  – длина;  $n_i$  – частота модального интервала, т.е. интервала, имеющего наибольшую частоту).

**3) Выборочная медиана  $\bar{M}_i$ :** медиана дискретного вариационного ряда при нечетном  $n$  определяется значением срединного элемента, при четном  $n$  равна среднему арифметическому двух срединных элементов; медиана

интервального вариационного ряда определяется по формуле

$$\overline{M}_l = x_0 + h \frac{n/2 - T_{i-1}}{n_i} \quad (\text{здесь } x_0 - \text{начало; } h - \text{длина; } n_i - \text{частота медианного}$$

интервала, т.е. интервала, содержащего срединный элемент;  $T_{i-1}$  – сумма частот интервалов, предшествующих медианному).

**Характеристики рассеяния** определяют разброс значений признака вокруг среднего значения  $\bar{x}$ .

**Выборочной дисперсией**  $D_B$  называют среднее арифметическое квадратов отклонений наблюдаемых значений признака от их среднего значения  $\bar{x}$ .

Выборочная дисперсия  $D_B = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$  – для несгруппированных данных и

$$D_B = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^2}{n} - \text{для сгруппированных данных.}$$

**Замечание.** Для малых выборок более точной (несмещенной) оценкой дисперсии является величина  $S^2 = \frac{n}{n-1} D_B$ , которую называют **исправленной выборочной дисперсией**. Так как  $\lim_{n \rightarrow \infty} \frac{n}{n-1} = 1$ , с увеличением объема выборки  $n$  точечные оценки  $D_B$  и  $S^2$  принимают близкие значения.

**Выборочный коэффициент асимметрии** вычисляется по формуле

$As = \mu_3 / \hat{\sigma}^3$ , где  $\mu_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n}$  – центральный момент 3-го порядка. Асимметрия характеризует отклонение от симметричности распределения. Если  $As > 0$ , то говорят о правосторонней асимметрии распределения (вытянутость вправо). Если  $As < 0$ , то говорят о левосторонней асимметрии распределения (вытянутость влево).

**Выборочный коэффициент эксцесса** вычисляется по формуле

$E = \mu_4 / \hat{\sigma}^4 - 3$ , где  $\mu_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n}$  – центральный момент 4-го порядка. Эксцесс



характеризует отступление от нормального распределения: для генерального нормального распределения значение  $E$  равно 0.

## Обработка данных в Excel

### 1. Сортировка

Сортировка данных помогает быстро придавать данным удобную форму и лучше понимать их, организовывать и находить необходимую информацию и в итоге принимать более эффективные решения.

Для сортировки данных в Excel выберите столбец с цифровыми данными в диапазоне ячеек или убедитесь, что активная ячейка находится в столбце таблицы, который содержит цифровые данные. После перехода **Главная** → **Редактирование** → **Сортировка и фильтр** (рис. 1.1)

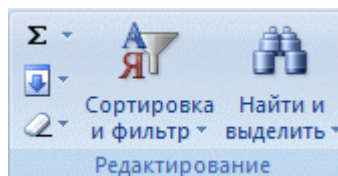


Рис. 1.1

выбираем одно из следующих действий (рис. 1.2):

- для сортировки чисел по возрастанию выберите вариант **Сортировка от минимального к максимальному**;
- для сортировки чисел по убыванию выберите вариант **Сортировка от максимального к минимальному**;
- для специализированной сортировки чисел выберите вариант **Настраиваемая сортировка**.

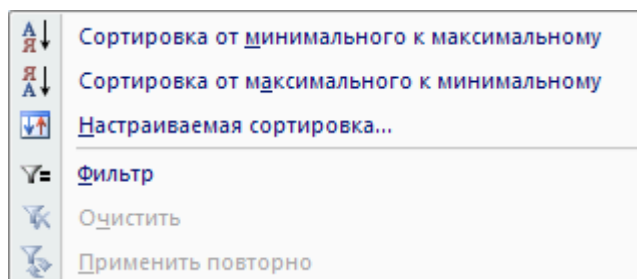


Рис. 1.2

Для сортировки в определенном пользователем порядке можно использовать пользовательские списки. В Excel предоставляются встроенные

пользовательские списки дней недели и месяцев года, однако также могут создаваться собственные пользовательские списки.

## 2. Группировка

Группировка элементов позволяет выделить набор данных, удовлетворяющих определенным требованиям, которые сложно выделить другим способом, например путем сортировки или фильтрации.

Для группировки числовых элементов в Excel выберите числовое поле, которое следует сгруппировать. Далее **Параметры** → **Группировать** → **Группировка по полю**, в итоге появится основной инструмент – диалоговое окно группировки данных (рис. 1.3)

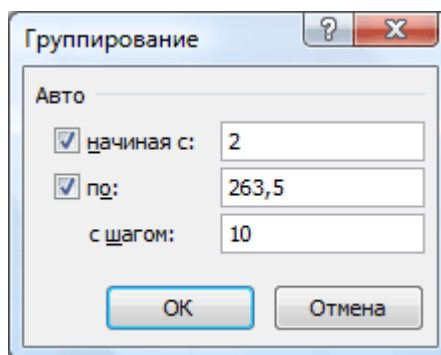


Рис. 1.3

В окне можно задать начальное и конечное значение интересующего интервала и шаг изменения.

## 3. Вычисление описательных статистик в Excel

Изучим различные способы вычисления описательных статистик: выборочное среднее, выборочную дисперсию, исправленную выборочную дисперсию, выборочное стандартное отклонение, асимметрию, моду, медиану.

**А) Вычисление описательных статистик по формулам.** Рассмотрим выборку – уровень безработицы в регионах Центральной России (данные взяты с сайта федеральной службы государственной статистики [www.gks.ru](http://www.gks.ru)). Так как количество наблюдений не велико, данные не будем группировать. Оценим выборочные СКО и показатель асимметрии распределения, осуществляя вычисления по формулам.

После ввода исходных данных  $n=10$  (столбцы А и В на рис. 1.4), вычисляем в столбце С выборочную среднюю  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i : = \text{СУММ(В2:В11)}/10$ .

Далее в столбцах D и E находим соответственно квадрат и куб отклонений вариант  $x_i$  от выборочного среднего  $\bar{x}$ . В связи с тем, что количество данных мало, будем находить выборочное СКО через исправленную выборочную дисперсию  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{n}{n-1} D_B$ , которая является несмещенной оценкой

генеральной дисперсии. Таким образом, в столбце F имеем:  $=10*\text{СУММ(D2:D11)}/9$ , после чего в столбце G находим интересующий нас показатель исправленного выборочного СКО  $S = \sqrt{\frac{n}{n-1} D_B} = \sqrt{S^2} : = \text{КОРЕНЬ(F2)}$ .

В столбцах H и I вычисляем центральный момент 3-го порядка  $\mu_3 : = \text{СУММ(E2:E11)}/10$  и показатель асимметрии, который с учётом поправки на несмещенность вычисляется как  $As = \frac{n^2}{(n-1)(n-2)} \frac{\mu_3}{S^3} : = 100*H2/(72*G2^3)$ .

	A	B	C	D	E	F	G	H	I
1		Ур. безработицы $x_i$ , %	$\bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^3$	$S^2$	$S$	$\mu_3$	$As$
2	Белгородская обл.	4,8	8,36	12,674	-45,118	4,523	2,127	-4,560	-0,658
3	Брянская обл.	10,7		5,476	12,813				
4	Воронежская обл.	8,6		0,058	0,014				
5	Ивановская обл.	10,8		5,954	14,527				
6	Курская обл.	8,8		0,194	0,085				
7	Липецкая обл.	5,6		7,618	-21,025				
8	Орловская обл.	9,9		2,372	3,652				
9	Рязанская обл.	9,2		0,706	0,593				
10	Тамбовская обл.	9,1		0,548	0,405				
11	Тульская обл.	6,1		5,108	-11,543				

Рис. 1.4

По полученным оценкам  $S$  и  $As$  можно сделать вывод о среднем отклонении показателя уровня безработицы в регионах Центральной России от величины  $\bar{x}$  и о тенденции вариантов генеральной совокупности к левостороннему смещению относительно выборочного среднего  $\bar{x}$ .

**Б) Вычисление описательных статистик с помощью статистических функций.** MS Excel предоставляет широкие возможности для анализа статистических данных. Для решения простых задач можно использовать встроенные статистические функции **СЧЕТ**, **МОДА**, **МЕДИАНА**, **СРЗНАЧ**,

**ДИСПР, ДИСП, СТАНДОТКЛОНП, СКОС и ЭКСЦЕСС**, которые позволяют определить для выборки объём, моду, медиану, выборочное среднее, выборочную дисперсию, исправленную выборочную дисперсию, выборочное стандартное отклонение, асимметрию и эксцесс соответственно. Для ознакомления со всеми статистическими функциями в Excel воспользуйтесь переходом: **Формулы** → **Библиотека функций** → **Другие функции** (рис. 1.5),

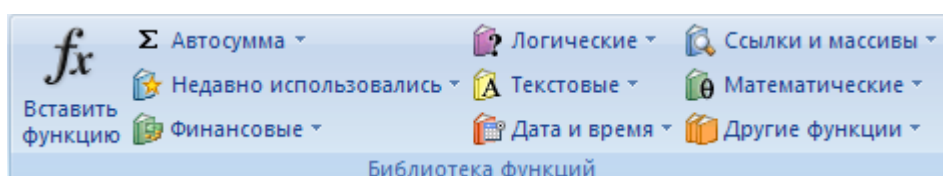


Рис. 1.5

далее из выпадающего списка выберите **Статистические** (рис. 1.6).

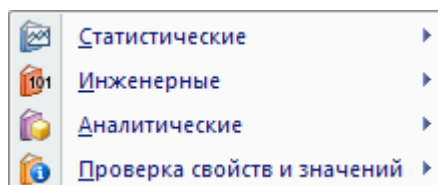


Рис. 1.6

### Задание для самостоятельной работы

Для закрепления навыков работы с Excel, вычислите с помощью статистических функций описательные статистики приведённой выше выборки и сравните результаты с полученными в пункте А данного раздела.

**В) Вычисление описательных статистик с помощью пакета *Анализ данных*.** Для проведения сложного статистического анализа можно упростить процесс и сэкономить время, используя надстройку «Пакет анализа» – дополнение Excel, расширяющее аналитические возможности и позволяющее строить гистограммы, составлять таблицы ранг и персентиль, делать случайные или периодические выборки данных и находить их статистические характеристики, генерировать неравномерно распределенные случайные числа, проводить регрессионный анализ и многое другое. Рассмотрим как с помощью «Пакета анализа» можно сразу получить все характеристики выборки.

Если данная надстройка ещё не подключена, последовательно переходим: **Параметры Excel** → **Надстройки** → **Перейти...** Далее среди доступных

настроек выбираем **Пакет анализа** и нажимаем **ОК**. Работа с установленной надстройкой происходит следующим образом: **Данные** → **Анализ данных** → **Описательная статистика** → **ОК** (рис. 1.7).

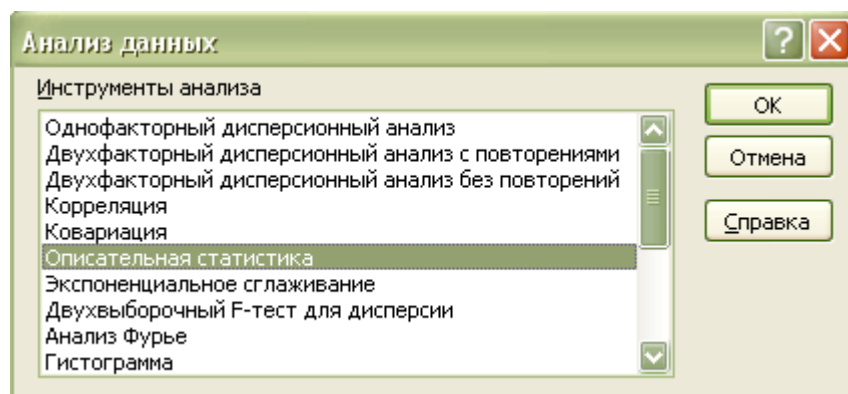


Рис. 1.7

Появляется диалоговое окно, которое требуется заполнить (рис. 1.8).

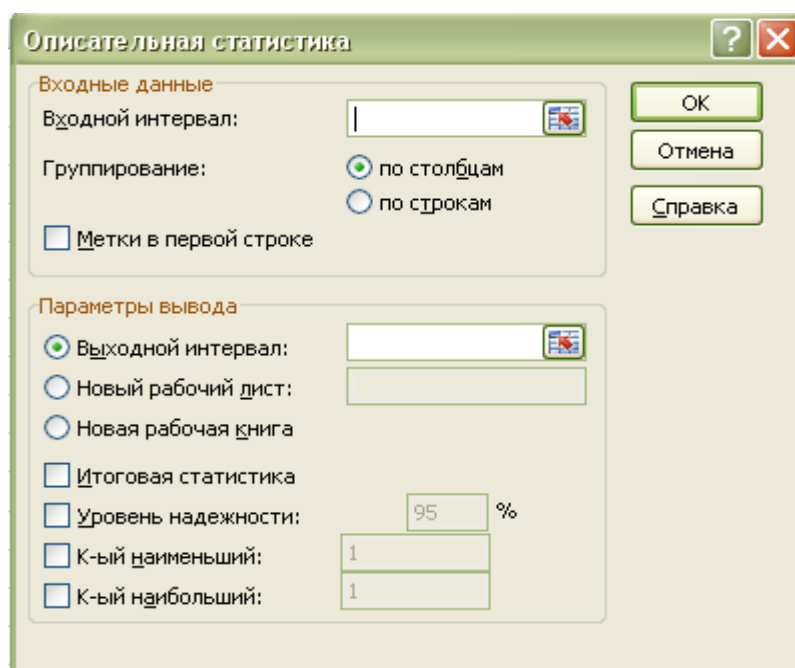


Рис. 1.8

В этом окне указываем, как сгруппированы данные, ставим «галочку» рядом со словами **Итоговая статистика**, нажимаем **ОК**. Появляется итоговое окно. Для рассмотренной выше выборки итоговый результат представлен на рис. 1.9.

	A	B	C	D
1		Ур. безрабо- тицы $x_i$ , %	Описательная статистика	
2	Белгородская обл.	4,8	Среднее	8,36
3	Брянская обл.	10,7	Стандартная ошибка	0,673
4	Воронежская обл.	8,6	Медиана	8,95
5	Ивановская обл.	10,8	Мода	#Н/Д
6	Курская обл.	8,8	Стандартное отклонение	2,127
7	Липецкая обл.	5,6	Дисперсия выборки	4,523
8	Орловская обл.	9,9	Эксцесс	-0,961
9	Рязанская обл.	9,2	Асимметричность	-0,658
10	Тамбовская обл.	9,1	Интервал	6
11	Тульская обл.	6,1	Минимум	4,8
12			Максимум	10,8
13			Сумма	83,6
14			Счет	10

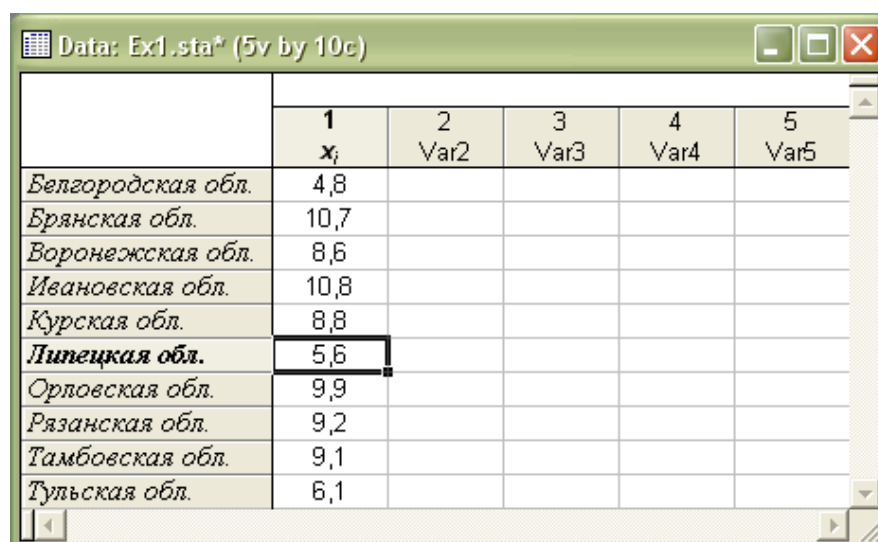
Рис. 1.9

### Вычисление описательных статистик в системе STATISTICA

STATISTICA — это универсальная интегрированная система, предназначенная для статистического анализа и обработки данных, которая содержит многофункциональную систему для работы с данными, широкий набор статистических модулей, в которых собраны группы логически связанных между собой статистических процедур, специальный инструментарий для подготовки отчетов, мощную графическую систему для визуализации данных, систему обмена данными с другими Windows-приложениями.

Вычисление описательных статистик в системе STATISTICA 6.0 можно произвести и непосредственно по формулам, и с использованием встроенных статистических функций по аналогии с Excel. Рассмотрим, как это можно сделать с помощью специализированного модуля *Descriptive statistics* (описательные статистики).

Сначала заносим в таблицу программы STATISTICA исследуемую выборку (рис. 1.10). Для запуска модуля *Descriptive statistics* осуществляем следующие переходы: *Статистика* → *Основная статистика/Таблицы* → *Descriptive statistics* → *OK*.



	1 $x_i$	2 Var2	3 Var3	4 Var4	5 Var5
Белгородская обл.	4,8				
Брянская обл.	10,7				
Воронежская обл.	8,6				
Ивановская обл.	10,8				
Курская обл.	8,8				
Липецкая обл.	5,6				
Орловская обл.	9,9				
Рязанская обл.	9,2				
Тамбовская обл.	9,1				
Тульская обл.	6,1				

Рис. 1.10

В появившемся меню (рис. 1.11) выбираем переменную, описательные статистики которой нас интересуют. Для выбора нажимаем кнопку **Variables** и в открывшемся окне щелкаем на имени переменной. Для просмотра результатов надо нажать кнопку **Summary: Descriptive statistics**. Откроется таблица с основными статистиками.

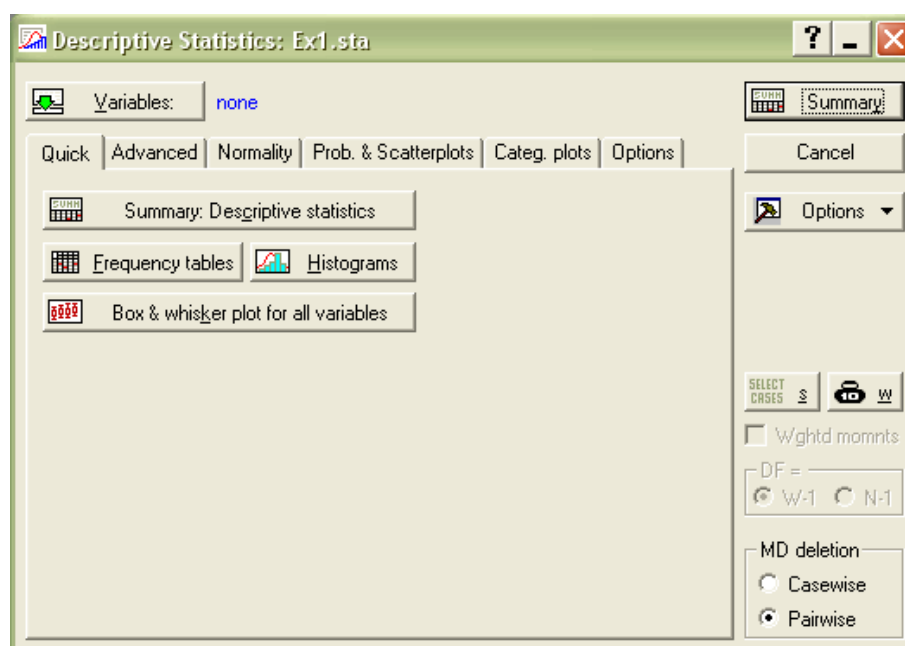


Рис. 1.11

Если нас интересуют другие статистики, необходимо указать их на вкладке **Advanced**, установив флажки напротив соответствующих статистик (рис. 1.12). При помощи кнопки **Select all stats** можно выбрать все статистики.

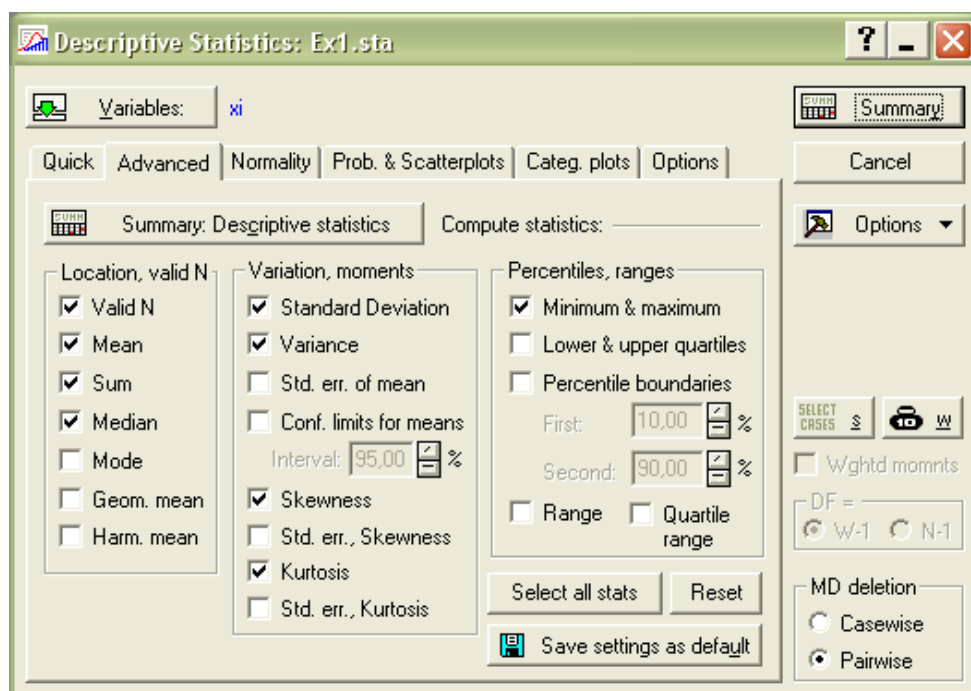


Рис. 1.12

Для исследуемой выборки выберем следующие статистики *Valid N* (объем выборки), *Mean* (выборочное среднее), *Sum* (сумма), *Median* (медиана), *Standard Deviation* (выборочное стандартное отклонение), *Variance* (выборочную дисперсию), *Skewness* (асимметрию), *Kurtosis* (эксцесс), *Minimum & maximum* (минимальная и максимальная варианты). Результирующая таблица представлена на рис. 1.13.

Variable	Valid N	Mean	Median	Sum	Minimum	Maximum	Variance	Std.Dev.	Skewness	Kurtosis
xi	10	8,360000	8,950000	83,60000	4,800000	10,80000	4,522667	2,126656	-0,658432	-0,960818

Рис. 1.13

Более подробно о возможностях модуля *Descriptive statistics* можно узнать в [14].

### Графический анализ данных. Гистограмма

Важным способом «описания» переменной является форма ее распределения, которая показывает, с какой частотой значения переменной попадают в определенные интервалы. Эти интервалы, называемые интервалами группировки, выбираются исследователем. Обычно исследователя интересует, насколько точно распределение можно аппроксимировать каким-либо



стандартным распределением, например нормальным. Простые описательные статистики дают об этом некоторую информацию. Более точную информацию о законе распределения можно получить с помощью графического анализа, а также с использованием специальных статистических критериев.

1. **Построение гистограммы в Excel.** Рассмотрим построение гистограммы в Excel с помощью «Пакета анализа»: *Данные* → *Анализ данных* → *Гистограмма* → *ОК*. После этого появится диалоговое окно (рис. 1.14), где в качестве входного интервала требуется ввести ссылку на ячейки, содержащие анализируемые данные. Установите флажок напротив слов **Вывод графика** и нажимаем *ОК*.

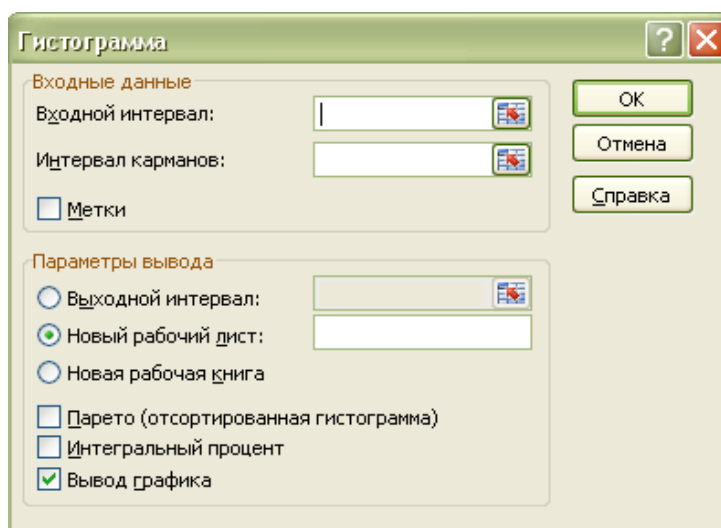


Рис. 1.14

Кроме этого пользователь может выбрать необязательный параметр **Интервал карманов** – набор граничных значений, определяющих отрезки (карманы). Excel вычисляет число попаданий данных в диапазон между текущим началом отрезка и соседним большим по порядку, если такой существует. Если диапазон карманов не введен, то набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

2. **Построение гистограммы в STATISTICA.** Рассмотрим построение гистограмм в системе STATISTICA с помощью модуля *Descriptive statistics*.

*Статистика* → *Основная статистика/Таблицы* → *Descriptive statistics* → *ОК*. Далее выбираем вкладку *Normality* (рис. 1.15), которая предназначена для

исследования возможности аппроксимации эмпирического закона распределения нормальным законом.

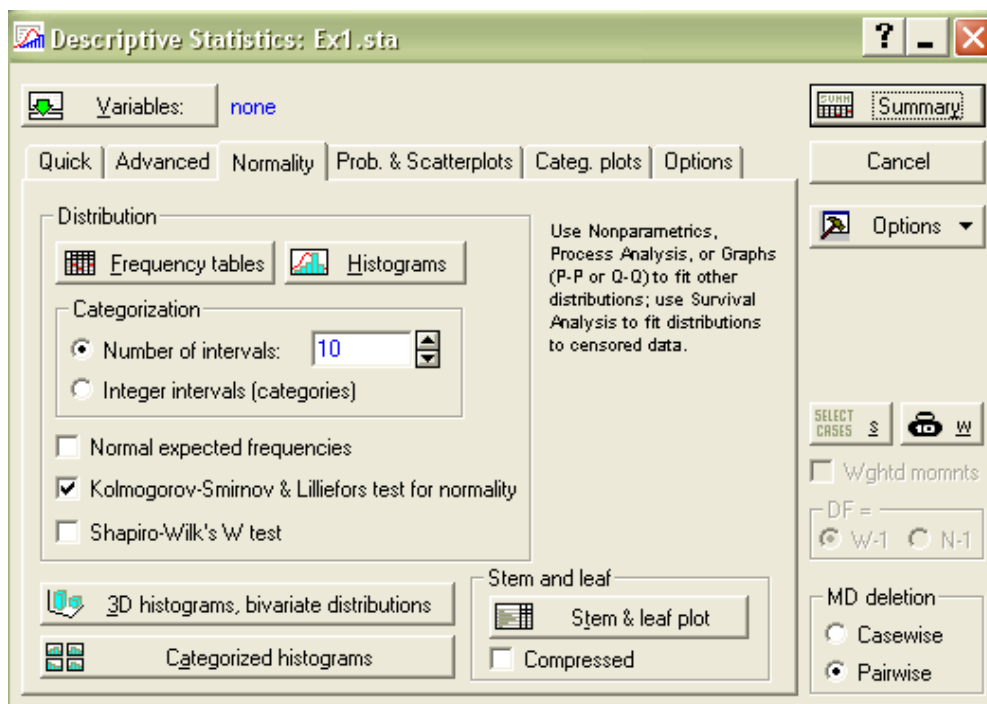


Рис. 1.15

Если установить флажок на *Number of intervals*, переменная воспринимается программой как непрерывная случайная величина и можно указать число интервалов разбиения диапазона ее изменения для построения гистограммы или *Frequency tables* (таблицы частот). При этом можно указать критерии соответствия эмпирического распределения нормальному закону (например, *Kolmogorov-Smirnov & Lilliefors test for normality*).

Если переменная является дискретной, то гистограмма визуализирует количественное соотношение различных значений переменной. Так, если установить флажок на *Integer intervals*, переменная воспринимается программой как дискретная случайная величина и число интервалов разбиения диапазона ее изменения определяется как число различных значений переменной.

После выбора всех параметров, предоставляемых программой, нажимаем кнопку *Histograms*. На рис. 1.16 показана гистограмма для показателя уровня безработицы.

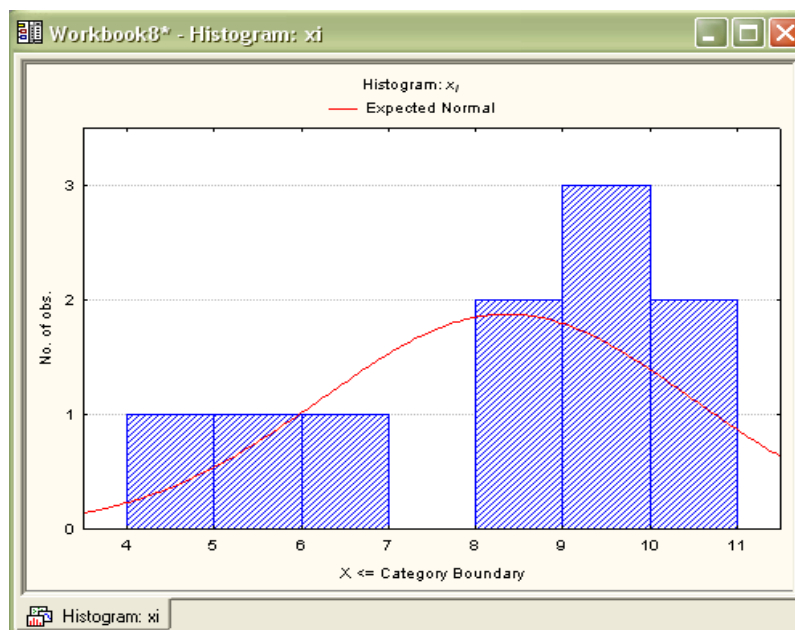


Рис. 1.16

Гистограмма показывает, что исследуемая выборка плохо аппроксимируется нормальным законом распределения.

### Задание для самостоятельной работы

1. Найти данные (в Интернете, журналах, статистических сборниках, справочниках). Используя средства Excel и STATISTICA, провести их группировку, графический анализ, вычисление и анализ описательных статистик.

2. В системе STATISTICA, используя возможности справки, рассмотреть различные способы графического представления данных.

### Задание для самостоятельного изучения

#### Генератор случайных чисел

Важно отметить, что компьютер можно использовать не только для автоматизации выполнения рутинных расчетов в ходе анализа и обработки данных, но и как инструмент познания при исследовании математических моделей.

#### Цели задания

**Знать** – определения понятий «случайное число» и «псевдослучайное число».

**Уметь** – генерировать последовательность случайных чисел на интервале  $[0;1)$  и осуществлять их отображение на произвольный интервал  $[a; b)$ .

**Владеть** – *навыками генерации случайных чисел в различных программных средах.*

**Статистическое моделирование** – построение математических имитаций случайных явлений или процессов. Это перспективное научное направление получило развитие в середине XX века в связи с ростом возможностей вычислительной техники и широко применяется для решения задач из различных областей человеческого знания. Например, расчеты систем массового обслуживания, расчеты качества и надежности изделий, задачи теории игр, задачи дискретной оптимизации, задачи финансовой математики, численное интегрирование, задачи динамики разреженного газа [1]. Появление методов статистического моделирования (Монте-Карло) в различных областях прикладной математики, как правило, связано с необходимостью решения качественно новых задач, возникающих из потребностей практики. Основа методов Монте-Карло – генератор случайных чисел (ГСЧ). О видах генераторов случайных чисел можно прочитать в [1,12].

Согласно [4], случайными числами будем называть возможные значения  $\xi_k$  равномерно распределенной случайной величины  $\xi$ , где  $0 \leq \xi < 1$ . Т.е. при генерировании последовательности случайных чисел  $\xi_1, \xi_2, \dots, \xi_n$  числовой промежуток  $[0; 1)$  покрывается равномерно: в интервалы равной длины попадает примерно одинаковое количество чисел.

Различают случайные числа, генерируемые каким-либо стохастическим устройством, и псевдослучайные числа, конструируемые с помощью арифметических алгоритмов.

### **Генератор случайных чисел в Excel**

Функция **СЛЧИС()** возвращает равномерно распределенное случайное число  $\xi$ , где  $0 \leq \xi < 1$ . Вместе с тем путем несложных преобразований с помощью функции **СЛЧИС()** можно получить любое случайное вещественное число. Например, чтобы получить случайное число между  $a$  и  $b$ , достаточно задать в любой ячейке таблицы Excel следующую формулу: **=СЛЧИС()\*(b-a)+a**.

Заметим, что начиная с Excel 2003 функция **СЛЧИС()** была улучшена. Теперь она реализует алгоритм Вичмана-Хилла, который проходит все

стандартные тесты на случайность и гарантирует, что повторение в комбинации случайных чисел начнётся не ранее, чем через  $10^{13}$  генерируемых чисел.

### Генератор случайных чисел в STATISTICA

Для генерации случайных чисел в STATISTICA надо дважды щелкнуть в таблице данных (в которой предполагается записать сгенерированные числа) на имени переменной. В окне спецификации переменной нажмите кнопку **Functions**. В открывшемся окне (рис. 1.17) надо выделить **Math** и выбрать функцию **Rnd**.

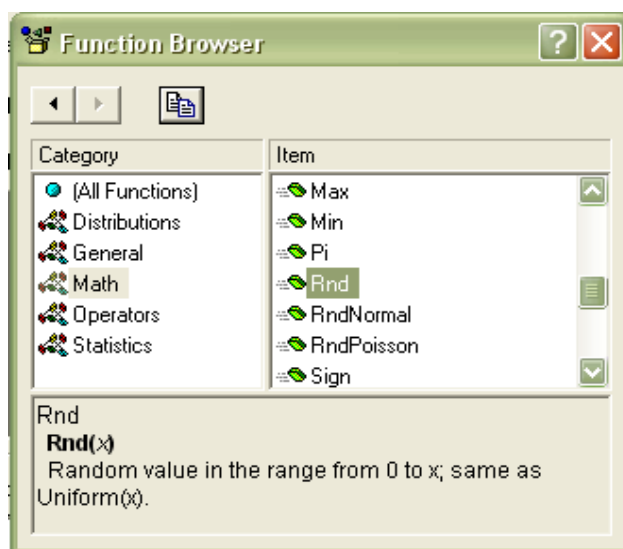


Рис. 1.17

**RND(X)** – генерация равномерно распределенных чисел. Эта функция имеет только один параметр – **X**, который задает правую границу интервала, содержащего случайные числа. При этом 0 является левой границей. Чтобы вписать общий вид функции **RND(X)** в окно спецификации переменной, достаточно дважды щелкнуть на имени функции в окне **Function Browser**. После указания числового значения параметра **X** надо нажать **OK**. Программа выдаст сообщение о правильности написания функции и запросит подтверждение о пересчете значения переменной. После подтверждения соответствующий столбец заполняется случайными числами.

### Задание для самостоятельной работы

1. Сгенерировать ряды из 10, 25, 50, 100 случайных чисел.
2. Вычислить описательные статистики
3. Построить гистограммы.

Какие выводы можно сделать относительно вида распределения? Будет ли оно равномерным? Как влияет количество наблюдений на данный вывод?

## Занятие 2

### Вероятность. Моделирование полной группы событий

#### Лабораторная работа № 1

Лабораторная работа представляет собой самостоятельное исследование с последующей защитой.

#### Цели занятия

- *Формирование навыков стохастического моделирования.*
- *Уяснение сущности и связи понятий «вероятность», «относительная частота», «статистическое определение вероятности».*
- *Экспериментальная проверка свойств вероятности и возможности вычисления вероятности случайного события опытным путем.*
- *Формирование навыков исследования явлений, имеющих вероятностную природу.*

Наблюдаемые нами события (явления) можно подразделить на следующие три вида: достоверные, невозможные и случайные.

**Достоверным** называют событие, которое обязательно произойдет, если будет осуществлена определенная совокупность условий  $S$ .

**Невозможным** называют событие, которое заведомо не произойдет, если будет осуществлена совокупность условий  $S$ .

**Случайным** называют событие, которое при осуществлении совокупности условий  $S$  может либо произойти, либо не произойти.

**Предметом теории вероятностей** является изучение вероятностных закономерностей массовых однородных случайных событий.

События называют **несовместными**, если появление одного из них исключает появление других событий в одном и том же испытании.

Несколько событий образуют **полную группу**, если в результате испытания появится хотя бы одно из них. Другими словами, появление хотя бы одного из событий полной группы есть достоверное событие.

События называют **равновозможными**, если есть основания считать, что ни одно из этих событий не является более возможным, чем другие.

Каждый из равновозможных результатов испытания называется **элементарным исходом**.

**Классическое определение вероятности:** вероятностью события  $A$  называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу.

Таким образом, вероятность события  $A$  определяется формулой  $P(A) = \frac{m}{n}$ , где  $m$  – число элементарных исходов, благоприятствующих событию  $A$ ,  $n$  – число всех возможных элементарных исходов испытания.

Одним из недостатков классического определения вероятности является то, что оно неприменимо к испытаниям с бесконечным числом исходов.

**Геометрическое определение** вероятности обобщает классическое на случай бесконечного числа элементарных исходов и представляет собой вероятность попадания точки в область (отрезок, часть плоскости и т.д.).

Таким образом, вероятность события  $A$  определяется формулой  $P(A) = \frac{\mu(A)}{\mu(\Omega)}$ , где  $\mu(A)$  – мера множества  $A$  (длина, площадь, объем);  $\mu(\Omega)$  – мера пространства элементарных событий.

Относительная частота, наряду с вероятностью, принадлежит к основным понятиям теории вероятностей.

**Относительной частотой события** называют отношение числа испытаний, в которых событие появилось, к общему числу фактически произведенных испытаний.

Таким образом, относительная частота события  $A$  определяется формулой  $W(A) = \frac{m}{n}$ , где  $m$  – число появлений события,  $n$  – общее число испытаний.

Еще одним недостатком классического определения вероятности следует считать то, что трудно указать основания, позволяющие считать элементарные события равновозможными. По этой причине наряду с классическим определением пользуются также **статистическим определением вероятности**, принимая за вероятность события относительную частоту или число, близкое к ней.

### **1. Моделирование случайного события, имеющего вероятность $p$ .**

Генерируется случайное число  $y$ , равномерно распределенное на отрезке  $[0; 1]$ . Если  $y \leq p$ , то событие  $A$  наступило.

### **2. Моделирование полной группы событий.**

Занумеруем события, образующие полную группу, числами от 1 до  $n$  (где  $n$  – количество событий) и составим таблицу: в первой строке – номер события, во второй – вероятность появления события с указанным номером.

Номер события	1	2	...	$j$	...	$n$
Вероятность события	$p_1$	$p_2$	...	$p_j$	...	$p_n$

Разобьем отрезок  $[0; 1]$  на оси  $Oy$  точками с координатами  $p_1, p_1+p_2, p_1+p_2+p_3, \dots, p_1+p_2+\dots+p_{n-1}$  на  $n$  частичных интервалов  $\Delta_1, \Delta_2, \dots, \Delta_n$ . При этом длина частичного интервала с номером  $j$  равна вероятности  $p_j$ .

Генерируется случайное число  $u$ , равномерно распределенное на отрезке  $[0; 1]$ . Если  $u$  принадлежит интервалу  $\Delta_j$ , то событие  $A_j$  наступило.

### **Лабораторная работа № 1. Экспериментальное вычисление вероятности.**

**Цели работы:** моделирование случайных событий, изучение свойств статистической вероятности события в зависимости от количества испытаний.

Лабораторную работу проведем в два этапа.

#### **Этап 1. Моделирование подбрасывания симметричной монеты.**

Событие  $A$  состоит в выпадении герба. Вероятность  $p$  события  $A$  равна 0,5.

а) Требуется выяснить, каким должно быть количество испытаний  $n$ , чтобы с вероятностью 0,9 отклонение (по абсолютной величине) относительной частоты появления герба  $m/n$  от вероятности  $p = 0,5$  не превышало числа  $\varepsilon > 0$ :  $P(|(m/n) - p| < \varepsilon) = 0,9$ .

Расчеты провести для  $\varepsilon = 0,05$  и  $\varepsilon = 0,01$ . Для вычислений воспользуемся следствием из интегральной теоремы Муавра-Лапласа:

$$P(|(m/n) - p| < \varepsilon) \approx 2\Phi(\varepsilon\sqrt{n/pq}), \text{ где } \Phi(x) = \int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt; \quad q=1-p.$$

Как связаны между собой значения  $\varepsilon$  и  $n$ ?

б) Провести  $k = 10$  серий по  $n$  испытаний в каждой. В скольких сериях неравенство  $|(m/n) - p| < \varepsilon$  выполнено и в скольких нарушено? Каким будет результат, если  $k \rightarrow \infty$ ?

#### **Этап 2. Моделирование реализации исходов случайного эксперимента.**

а) Разработать алгоритм моделирования реализации опыта со случайными исходами согласно индивидуальным заданиям (см. прил. 1).



б) Разработать программу (программы) для моделирования реализации исходов опыта определённое конечное число раз, с обязательным сохранением начальных условий опыта и для расчёта частоты появления интересующего события.

в) Составить статистическую таблицу зависимости частоты появления заданного события от числа проведённых опытов.

г) По статистической таблице построить график зависимости частоты события от числа опытов.

д) Составить статистическую таблицу отклонений значений частоты события от вероятности появления этого события.

е) Отобразить полученные табличные данные на графиках.

ж) Найти значение  $n$  (число испытаний), чтобы  $P(|(m/n) - p| < 0,05) = 0,9$  и  $P(|(m/n) - p| < 0,01) = 0,9$ .

Сделать выводы по работе.

### Занятие 3

#### Моделирование случайной величины с заданным распределением

Моделирование (разыгрывание) случайной величины с заданным распределением – процесс вычисления последовательности ее возможных значений. Методы моделирования случайных величин рассмотрены в работах [1,4–6, 9, 11, 12].

#### Цели занятия

**Знать** – способы задания, параметры и основные алгоритмы моделирования дискретных и непрерывных случайных величин.

**Уметь** – генерировать последовательность значений случайной величины с заданным законом распределения.

**Владеть** – различными методами разыгрывания (моделирования) случайной величины с заданным распределением, используя ЭВМ (программная реализация алгоритмов, возможности Excel и STATISTICA).

#### 1. Моделирование дискретной случайной величины

Пусть дискретная случайная величина  $X$  задана законом распределения

$X$	$x_0$	$x_1$	$\dots$	$x_k$	$\dots$
$P$	$p_0$	$p_1$	$\dots$	$p_k$	$\dots$

,

где  $P(X = x_k) = p_k$ ;  $\sum_{k=0} p_k = 1$ ;  $k = 0, 1, 2, \dots$

Рассмотрим стандартный алгоритм определения значения разыгрываемой дискретной случайной величины.

*Шаг 1.* Генерируем случайное число  $y$ , равномерно распределенное на отрезке  $[0; 1]$ ; полагаем  $B = y$  и  $k = 0$ .

*Шаг 2.* Полагаем  $B = B - p_k$ .

*Шаг 3.* Если  $B > 0$ , то увеличиваем значение  $k$  на 1 (т.е. полагаем  $k = k+1$ ) и переходим к шагу 2. Если  $B \leq 0$ , то переходим к шагу 4.

*Шаг 4.* Случайная величина  $X$  приняла значение  $x_k$ :  $X = x_k$ . Реализация значения случайной величины  $X$  получена.

**Замечание.** Моделирование дискретной случайной величины с конечным числом значений аналогично моделированию полной группы событий.

Среди дискретных распределений наиболее важными являются биномиальное распределение, распределение Пуассона, геометрическое и гипергеометрическое. Важно отметить, что случайные величины, распределенные по данным законам, принимают целые значения (т.е.  $X = k$ ,  $k = 0, 1, 2, \dots$ ), а вероятности  $p_k = P(X = k)$  связаны между собой рекуррентными соотношениями

$$p_{k+1} = r(k) \cdot p_k.$$

Для таких случайных величин нет необходимости хранить в памяти ЭВМ значения  $p_k$  и  $x_k$ . Их моделирование осуществляется по следующему алгоритму.

*Шаг 1.* Генерируем случайное число  $y$ , равномерно распределенное на отрезке  $[0; 1]$ ; полагаем  $B = y$ ,  $P = p_0$  и  $k = 0$ .

*Шаг 2.* Полагаем  $B = B - P$ .

*Шаг 3.* Если  $B \geq 0$ , то увеличиваем значение  $k$  на 1 (т.е.  $k = k+1$ ), вычисляем новое значение переменной  $P$  по формуле  $P = r(k) \cdot P$  и переходим к шагу 2. Если  $B < 0$ , то переходим к шагу 4.

*Шаг 4.* Случайная величина  $X$  приняла значение, равное  $k$ :  $X = k$ . Реализация значения случайной величины  $X$  получена.

### **Биномиальное распределение**

Говорят, что случайная величина  $X$  имеет **биномиальное** распределение с параметрами  $n$  и  $p$ , если  $X$  принимает значения  $0, 1, 2, \dots, n$  с вероятностями

$$p_k = P(X = k) = C_n^k p^k q^{n-k},$$

где  $X$  – число успехов в  $n$  независимых испытаниях;  $p$  – вероятность успеха в одном испытании;  $q = 1 - p$ . Тогда  $r(k) = \frac{p_{k+1}}{p_k} = \frac{n-k}{k+1} \cdot \frac{p}{q}$  при  $k = 0, 1, \dots, n$ .

### Распределение Пуассона

Говорят, что случайная величина  $X$  имеет *распределение Пуассона* с параметром  $\lambda > 0$ , если  $X$  принимает значения  $0, 1, 2, 3, \dots$  с вероятностями  $p_k = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ . Например,  $X$  – число независимых событий в фиксированный промежуток времени. Тогда  $r(k) = \frac{p_{k+1}}{p_k} = \frac{\lambda}{k+1}$ ,  $k=0, 1, 2, \dots$

### Геометрическое распределение

Говорят, что случайная величина  $X$  имеет *геометрическое* распределение с параметром  $p$ , если  $X$  принимает значения  $0, 1, 2, 3, \dots$  с вероятностями  $p_k = P(X = k) = pq^k$ , где  $q = 1 - p$ . Например, проводится серия из независимых испытаний,  $p$  – вероятность успеха в одном испытании. Случайная величина  $X$  – число испытаний до первого успеха, имеет геометрическое распределение. Тогда  $r(k) = \frac{p_{k+1}}{p_k} = q$ ,  $k = 0, 1, 2, \dots$

### Гипергеометрическое распределение

Говорят, что случайная величина  $X$  имеет *гипергеометрическое* распределение с параметрами  $n, N$  и  $K$  ( $n \leq N, K \leq N$ ), если  $X$  принимает целые значения от  $\max(0; N - K - n)$  до  $\min(k; n)$  с вероятностями  $p_k = P(X = k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}$ .

Например, пусть из совокупности, в которой имеется  $N$  шаров, причем из них  $K$  белых, извлекли  $n$  шаров. Тогда случайная величина  $X$  (число белых шаров в выборке) имеет гипергеометрическое распределение. В этом случае коэффициент  $r(k) = \frac{p_{k+1}}{p_k} = \frac{(K-k)(n-k)}{(k+1)(N-K-n+1+k)}$ ,  $\max(0; n+K-N) \leq k \leq \min(k; n)$ .

## 2. Моделирование непрерывной случайной величины

### Метод обратных функций

**Правило 1.** Пусть известна функция распределения  $F(x)$  непрерывной случайной величины  $X$ . Требуется разыграть  $X$ , то есть найти последовательность ее реализаций  $x_k$  ( $k=1, 2, \dots$ ). Вычисления производим по следующему алгоритму.

*Шаг 1.* Генерируем последовательность равномерно распределенных на отрезке  $[0; 1]$  случайных чисел  $y_1, y_2, \dots, y_k, \dots$

*Шаг 2.* Решаем относительно  $x_k$  уравнения  $F(x_k) = y_k$ . Последовательность чисел  $x_1 = F^{-1}(y_1), x_2 = F^{-1}(y_2), \dots, x_k = F^{-1}(y_k), \dots$  – реализация непрерывной случайной величины  $X$  – с функцией распределения  $F(x)$ .

**Правило 2.** Пусть известна функция плотности  $f(x)$  непрерывной случайной величины  $X$ . Требуется разыграть  $X$ , то есть найти последовательность ее реализаций  $x_k$  ( $k = 1, 2, \dots$ ). Вычисления производим по следующему алгоритму.

*Шаг 1.* Генерируем последовательность равномерно распределенных на отрезке  $[0; 1]$  случайных чисел  $y_1, y_2, \dots, y_k, \dots$

*Шаг 2.* Решаем относительно  $x_k$  уравнения  $\int_{-\infty}^{x_k} f(x)dx = y_k$ . Последовательность чисел  $x_k$  ( $k = 1, 2, \dots$ ) – реализация непрерывной случайной величины  $X$  – с функцией плотности  $f(x)$ .

**Пример 3.1.** Пусть случайная величина  $X$  имеет экспоненциальное (показательное) распределение с параметром  $\lambda > 0$ . Тогда функция распределения имеет вид  $F(x) = \begin{cases} 0, & x < 0; \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$

Требуется разыграть  $X$ , используя метод обратных функций.

В соответствии с правилом 1 генерируем последовательность значений  $y_1, y_2, \dots, y_k$ , равномерно распределенных на отрезке  $[0; 1]$ . Далее решаем уравнения  $F(x_k) = y_k$ , т. е.  $1 - e^{-\lambda x_k} = y_k$ . Получаем, что  $x_k = -\frac{\ln(1 - y_k)}{\lambda}$ .

Если  $y_k$  – случайное число из отрезка  $[0; 1]$ , то  $1 - y_k$  также является случайным числом из  $[0; 1]$ .

Тогда  $x_1 = -\frac{\ln y_1}{\lambda}, x_2 = -\frac{\ln y_2}{\lambda}, \dots, x_k = -\frac{\ln(y_k)}{\lambda}$  – значения случайной величины  $X$ , распределенной по экспоненциальному (показательному) закону с параметром  $\lambda$ . Таким образом, для разыгрывания случайной величины  $X$  найдена явная формула вида  $x = F^{-1}(y)$ .

**Замечание.** Получить решения уравнений  $y = F(x)$  в явном виде удастся не всегда. Например, в случае нормального распределения случайной величины  $X$ . В этом случае для нахождения решений применяют численные методы решения уравнений или используют другие методы для генерирования случайных величин [5, 9, 11, 12].

### **Моделирование случайной величины с нормальным распределением**

Так как метод обратных функций является неэффективным в случае моделирования нормально распределенных случайных величин, рассмотрим приближенный метод, основанный на центральной предельной теореме.

*Шаг 1.* Генерируем 12 значений случайной величины, имеющей равномерное распределение на отрезке  $[0; 1]$ :  $y_1, y_2, \dots, y_{12}$ .

*Шаг 2.* Возможное значение случайной величины  $X$ , имеющей нормальное распределение с параметрами  $a = 0$  и  $\sigma = 1$ , вычисляется по формуле

$$x_j = \sum_{k=1}^{12} y_k - 6.$$

Для получения следующего значения случайной величины  $X$  необходимо сгенерировать другие 12 значений  $y_1, y_2, \dots, y_{12}$  и т.д.

Если требуется разыграть нормально распределенную случайную величину  $Z$  с параметрами  $a$  и  $\sigma$ , то дополнительно выполняем шаг 3.

*Шаг 3.* Вычисляем  $z_j = a + \sigma \cdot x_j$ .

### **Задание для самостоятельной работы**

Найдите явную формулу для разыгрывания случайной величины  $X$  с помощью метода обратных функций, если

- a)  $X$  имеет равномерное распределение на отрезке  $[a; b]$ ;
- b)  $X$  распределена по закону Вейбулла с функцией плотности

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{n}{b} x^{n-1} e^{-x^n/b}, & x \geq 0; \end{cases}$$

- c)  $X$  распределена по закону Релея с функцией плотности

$$f(x) = \begin{cases} 0, & x < 0; \\ \frac{x}{\sigma^2} e^{-x^2/2\sigma^2}, & x \geq 0; \end{cases}$$

- d)  $X$  распределена по закону Коши с функцией плотности  $f(x) = \frac{1}{\pi} \cdot \frac{1}{x^2 + 1}$ ;

е)  $X$  задана функцией плотности  $f(x) = \begin{cases} 0, & x \notin [0;2]; \\ 1 - \frac{x}{2}, & x \in [0;2]. \end{cases}$

**Указание.** Найти функцию распределения  $F(x)$ , а затем решить относительно  $x$  уравнение  $F(x) = y$ .

**Замечание 1.** Считают, что время безотказной работы многих технических устройств является случайной величиной, имеющей распределение Вейбулла.

**Замечание 2.** Если в распределении Вейбулла параметр  $n = 1$ , то получим экспоненциальное (показательное) распределение, а если  $n = 2$ , то распределение Релея.

### 3. Моделирование случайных величин в Excel

**А) С помощью статистических функций.** В основе данного способа моделирования непрерывных случайных величин лежит метод обратных функций.

Пусть известна явная формула разыгрывания случайной величины  $X$  с заданным распределением  $x = F^{-1}(y)$ , где  $F^{-1}(y)$  – функция, обратная функции распределения  $F(x)$ ;  $y$  – случайное число из отрезка  $[0;1]$ . Вводим данную формулу, используя **Мастер функций** Excel.

**Пример 3.2.** Требуется разыграть значение случайной величины  $X$ , имеющей экспоненциальное (показательное) распределение с параметром  $\lambda = 2$ . Воспользуемся явной формулой, полученной в предыдущем примере:  $x = -\frac{\ln(y)}{\lambda}$ . Получаем значение случайной величины с помощью формулы: **=-LN(СЛЧИС())/2**.

Разыграть значения нормально распределенной случайной величины  $X$ , имеющей нормальное распределение с параметрами  $a$  и  $\sigma$ , можно с помощью статистической функции **НОРМОБР**, задав в качестве аргументов случайное число из отрезка  $[0;1]$ , значение математического ожидания  $a$  и значение среднеквадратического отклонения  $\sigma$ .

**Пример 3.3.** Требуется разыграть значение нормально распределенной случайной величины  $X$  с параметрами  $a = 2$  и  $\sigma = 1$ . Получаем результат с помощью формулы: **=НОРМОБР(СЛЧИС(); 2; 1)**.

**В) С помощью пакета Анализ данных.** Excel позволяет моделировать случайные как непрерывные, так и дискретные случайные величины.

Для моделирования случайной величины с заданным законом распределения осуществляем последовательность переходов: *Сервис* → *Анализ данных* → *Генерация случайных чисел* → *ОК*. Для Microsoft Office Excel 2007 последовательность имеет вид: *Данные* → *Анализ данных* → *Генерация случайных чисел* → *ОК* (рис. 3.1).

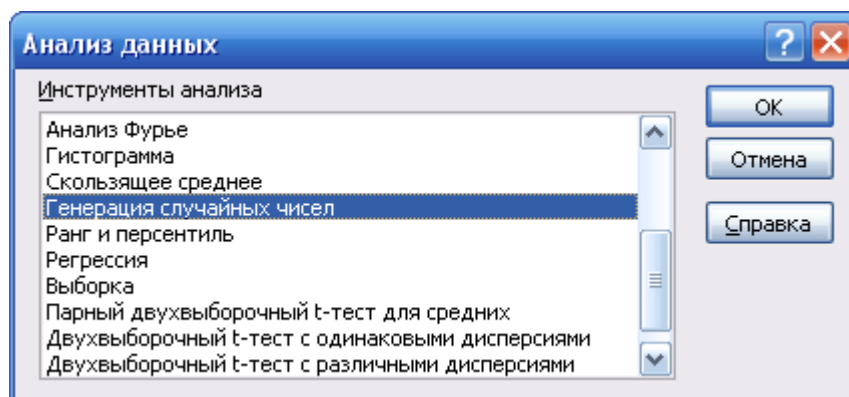


Рис. 3.1

В результате откроется диалоговое окно, в котором необходимо заполнить поля. *Число переменных* – количество столбцов значений, которое будет размещено в выходном диапазоне. *Число случайных чисел* – количество случайных чисел, сгенерированных в каждом столбце. *Распределение* – вид распределения (*Равномерное, Нормальное, Биномиальное, Дискретное, Бернулли, Пуассона, Модельное*). Для каждого вида распределения необходимо указать параметры. Например, параметрами равномерного распределения являются границы отрезка  $[a; b]$ , параметрами нормального распределения являются значение математического ожидания  $\mu$  и значение среднеквадратического отклонения  $\sigma$ , параметрами биномиального распределения являются количество независимых испытаний  $n$  и  $p$  – значение вероятности появления события в одном испытании, распределение Пуассона имеет один параметр  $\lambda$ , ( $\lambda > 0$ ). Кроме того, необходимо указать параметры вывода (*Выходной интервал, Новый рабочий лист, Новая рабочая книга*). После заполнения полей нажимаем *ОК*.

**Пример 3.4.** Требуется сгенерировать два столбца по сто значений в каждом, разыграв случайную величину  $X$ , имеющую равномерное распределение на отрезке  $[1; 3]$ . Результаты разместить на новом рабочем листе. Вводим значения, как показано на рисунке (см. рис. 3.2).

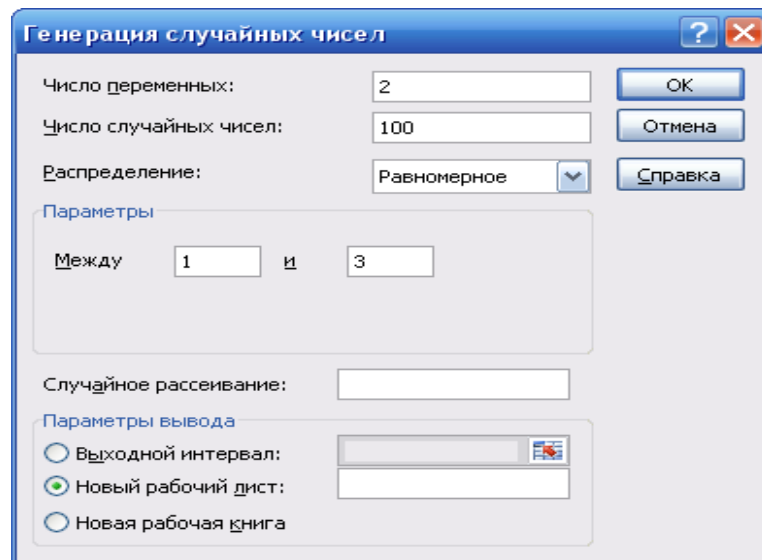


Рис 3.2

### Задание для самостоятельной работы

1. Используя явные формулы для разыгрывания случайных величин, смоделируйте с помощью **Мастера функций** Excel по 100 значений для каждой из случайных величин, рассмотренных в задании на с. 28.

2. Используя пакет *Анализ данных*, смоделируйте 2 столбца по 100 значений в каждом для случайных величин, имеющих следующие распределения: равномерное, нормальное, биномиальное, Пуассона. Постройте гистограммы для полученных выборок. Параметры распределений указаны в индивидуальном домашнем задании или задаются преподавателем.

### 4. Моделирование случайных величин в системе STATISTICA

Напомним, что на занятии 1 была рассмотрена функция  $RND(X)$  (генерация равномерно распределенных чисел на отрезке  $[0; X]$ ). Рассмотрим возможности разыгрывания случайных величин с другими непрерывными законами распределения.

Прежде всего необходимо создать файл данных и двойным щелчком правой кнопки мыши на поле имени переменной открыть окно спецификации переменной. Если по методу обратных функций получена явная формула разыгрывания случайной величины с заданным распределением, вводим ее в поле **Long name**, как, например, введена формула для разыгрывания случайной величины  $X$ , имеющей экспоненциальное распределение с параметром  $\lambda = 2$  (см. рис. 3.3).



**Variable 2**

Name:  MD code:

Display Format

Column width:  Decimals:

Category:	Representation:
Number	1 000,000; -1 000,000
Date	1 000,000; -1 000,000
Time	1 000,000; (1 000,000)
Scientific	1 000,000; (1 000,000)
Currency	
Percentage	

Long name (label, link, or formula with Functions):

Examples: Label: Gross income in 1991 Formulas: = v1 + v2 ; comment  
Link: @Excel[c:\file.xls!r2c2:r4c4] = (v1>0)\*AGE + v3

Рис. 3.3

Для ряда распределений STATISTICA позволяет вычислить значение функции  $F^{-1}(y)$ , обратной к функции распределения  $F(x)$ . Например, для того, чтобы смоделировать нормально распределенную случайную величину  $X$  с параметрами  $\mu = 2$  и  $\sigma = 1$ , в поле **Long name** окна спецификации переменной вводим формулу: `=VNormal(Rnd(1); 2; 1)` (рис. 3.4). Сравните с аналогичной функцией **НОРМОБР** Мастера функций Excel.

**Variable 1**

Name:  MD code:

Display Format

Column width:  Decimals:

Category:	Representation:
Number	1 000,000; -1 000,000
Date	1 000,000; -1 000,000
Time	1 000,000; (1 000,000)
Scientific	1 000,000; (1 000,000)
Currency	
Percentage	

Long name (label, link, or formula with Functions):

Examples: Label: Gross income in 1991 Formulas: = v1 + v2 ; comment  
Link: @Excel[c:\file.xls!r2c2:r4c4] = (v1>0)\*AGE + v3

Рис. 3.4

**Замечание.** Аналогичным образом с помощью обратных функций **VExpon**, **VRayl**, **VWeibull** могут быть разыграны случайные величины,

имеющие экспоненциальное распределение, распределение Релея, распределение Вейбулла. Нажав в поле **Long name** окна спецификации переменной кнопку **Functions**, и выбрав в открывшемся окне категорию **Distributions**, можно увидеть список распределений, для которых реализованы обратные функции (рис. 3.5).

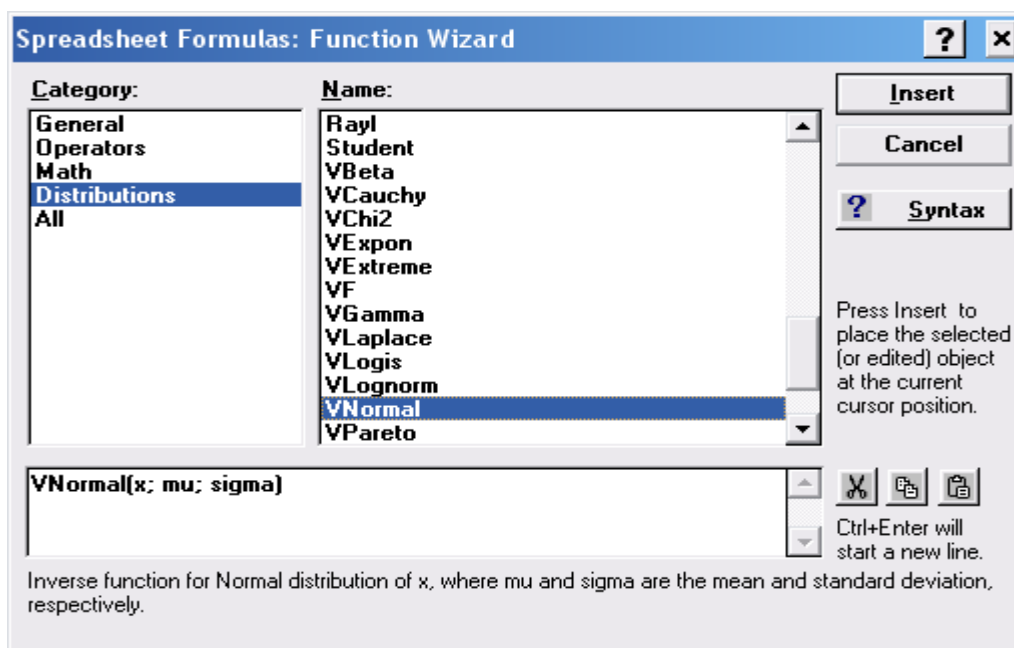


Рис. 3.5

Далее, выбрав в списке функций нужную нам функцию и нажав кнопку **Syntax**, можно изучить синтаксис и получить информацию о параметрах распределения, а нажав кнопку **Insert**, можно вставить заголовок и список параметров функции в формулу поля **Long name** окна спецификации переменной.

Приведем пример генерирования дискретной случайной величины. Пусть случайная величина  $X$  принимает два значения: 0 и 1. Причем  $X = 1$  с вероятностью  $p = 0,4$ . Требуется разыграть данную случайную величину. Результатом будет последовательность 0 и 1. Двойным щелчком в поле имени переменной открываем окно спецификации переменной, в поле **Long name** вводим формулу: **=(Rnd(1)<=0,4)\*1+0** и нажимаем **OK**.

Если требуется разыграть случайную величину, имеющую биномиальное распределение с параметрами  $n$  и  $p$ , то указанным выше способом генерируем  $n$  столбцов, содержащих последовательности из 0 и 1, а затем суммируем полученные результаты. Сумма будет представлять собой результат моделирования случайной величины, распределенной по биномиальному закону.

Также в программе STATISTICA 6.0 имеется возможность генерировать случайные числа, подчиняющиеся нормальному (*RndNormal*) и пуассоновскому (*RndPoisson*) законам распределения.

### **Задание для самостоятельной работы**

Смоделируйте 2 столбца по 100 значений в каждом для случайных величин, имеющих следующие распределения: равномерное, нормальное, биномиальное, экспоненциальное, Релея, Вейбулла.

Постройте гистограммы для полученных выборок. Параметры распределений указаны в индивидуальном домашнем задании или задаются преподавателем.

## **Занятие 4**

### **Вычисление определённого интеграла методом Монте-Карло**

#### **Лабораторная работа № 2**

Применение метода статистических испытаний для вычисления определённого интеграла рассмотрено в [6, 10, 12].

#### **Цели занятия**

- *Формирование навыков стохастического моделирования.*
- *Уяснение сущности и связи понятий «вероятность», «геометрическое определение вероятности», «сходимость по вероятности».*
- *Экспериментальная проверка возможности применения стохастического моделирования в ситуации, не имеющей вероятностной природы (численное интегрирование).*

**Методы Монте-Карло (методы статистических испытаний)** – это численные методы решения математических задач (систем алгебраических, дифференциальных, интегральных уравнений) и прямое статистическое моделирование (физических, химических, биологических, экономических, социальных процессов) при помощи получения и преобразования случайных чисел. То есть часть задач имеют очевидную вероятностную природу, а часть представляют собой пример применения идей статистического моделирования для исследования математических моделей объектов, не имеющих таковой (например, вычисление определённого интеграла).

Первая работа по использованию методов Монте-Карло была опубликована в 1873 году при организации стохастического процесса

экспериментального определения числа  $\pi$  путём бросания иглы на лист линованной бумаги. Своё романтическое название методы Монте-Карло получили по имени столицы княжества Монако, знаменитой своими игорными домами, основу которых составляет рулетка – совершенный инструмент для получения случайных чисел. А первая работа, где этот вопрос излагался систематически, опубликована в 1949 году Н. Метрополисом и С. Уламом, где метод Монте-Карло применялся для решения линейных интегральных уравнений, в котором явно угадывалась задача о прохождении нейтронов через вещество. В нашей стране работы по методам Монте-Карло стали активно публиковаться после Женевской международной конференции по применению атомной энергии в мирных целях.

## ***Лабораторная работа 2. Вычисление определённого интеграла методом Монте-Карло.***

***Цели работы:*** вычисление определённого интеграла от произвольной функции с помощью метода Монте-Карло с наперёд заданной точностью, определение числа испытаний, при котором относительная частота обладает свойством устойчивости для поставленной задачи.

Выполнение лабораторной работы разобьем на два этапа.

### ***Порядок выполнения работы на первом этапе***

Вычислить интеграл вида  $\int_0^1 x^k dx$  в явном виде, используя правила

вычисления определенного интеграла (значение  $k$  совпадает с номером в журнале или задается преподавателем).

1. Рассчитать  $n$  – необходимое количество испытаний, чтобы с вероятностью  $p = 0,9$  обеспечить точность вычисления интеграла  $\varepsilon = 0,01$ .
2. Вычислить определенный интеграл от функции  $f(x) = x^k$  на отрезке  $[0; 1]$ , смоделировав  $n$  испытаний.
3. Смоделировать 10 экспериментов по вычислению интеграла по  $n$  проб в каждом.
4. По результатам заполнить таблицу полученных результатов, сравнив значения интеграла, вычисленные по методу Монте-Карло, с точным значением определенного интеграла.

5. Сделать выводы о возможности применения метода статистических испытаний к решению задачи о численном нахождении определенного интеграла.

В качестве примера применим метод Монте-Карло для вычисления интеграла  $\int_0^1 x^2 dx$  в заштрихованной области (рис. 4.1).

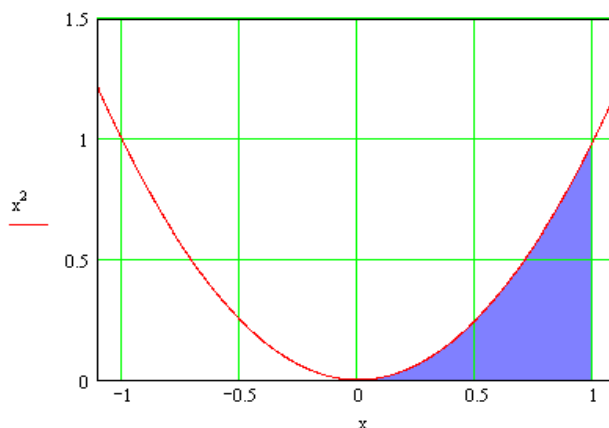


Рис. 4.1

а) Выберем в квадрате ( $a=1$  в первой четверти)  $n$  случайных точек. Пусть  $m$  точек попали в заданную область (ниже графика заданной функции). Рассмотрим отношение  $\frac{m}{n} \approx \frac{S_1}{S_2}$ , где  $S_1$  – площадь заданной фигуры;  $S_2=1$  – площадь квадрата (см. геометрическое определение вероятности).

Заметим, так как  $S_2=1$ , то интеграл  $\int_0^1 f(x)dx$  численно равен  $p$ . В нашем случае  $p = \int_0^1 x^2 dx = \frac{1}{3}$ .

б) Рассчитаем необходимое количество испытаний, чтобы получить значение интеграла с заданной точностью.

Пусть  $p = 0,5$ ,  $q = 1 - p = 0,5$ ,  $\varepsilon = 0,01$ ,  $P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 0,9$ . Подставим значения:

$$2\Phi\left(0,01 \cdot \sqrt{\frac{n}{0,5 \cdot 0,5}}\right) = 2\Phi(0,02 \cdot \sqrt{n}) = 0,9 \Rightarrow \Phi(0,02 \cdot \sqrt{n}) = 0,45.$$

Получим,  $0,02 \cdot \sqrt{n} = 1,65 \Rightarrow n = 6806$ .

в) Проведем 10 экспериментов  $s_i$  ( $1 \leq i \leq 10$ ) по  $n$  проб в каждом. С помощью программы получим результаты  $(x_i, y_i)$  ( $1 \leq i \leq n$ ), где  $x_i, y_i \in [0; 1]$ .

Пусть событие  $A$  – соответствует попаданию  $(x_i, y_i) \in [0; 1] \times [0; 1]$  в заштрихованную область,  $m$  – количество появлений события  $A$  в эксперименте.

Составим таблицу для  $n = 6806$ :

$s$	$s_1$	$s_2$	$s_3$	$s_4$	$s_5$	$s_6$	$s_7$	$s_8$	$s_9$	$s_{10}$
$m$										
$\frac{m}{n} = \int_0^1 f(x)dx$										
$\left  p - \frac{m}{n} \right $										

г) Сделаем выводы относительно значений отклонения  $|p - m/n|$  в 10 проведенных экспериментах. Обладает ли относительная частота свойством устойчивости? Что означает понятие «сходимость по вероятности» и подтверждает ли серия проведенных экспериментов тот факт, что относительная частота сходится по вероятности к теоретическому значению вероятности? Какое определение вероятности лежит в основе статистического метода вычисления значения определенного интеграла? Что способствовало возможности посредством применения статистических испытаний решить задачу вычисления определенного интеграла, которая не имеет вероятностной природы?

### ***Порядок выполнения работы на втором этапе***

1. Подобрать сложную функцию (желательно не имеющей первообразной в элементарных функциях), удовлетворяющую условиям варианта (см. прил. 2), и построить её график.

2. Правильно определить область для генерации пар случайных чисел.

3. Разработать программу, вычисляющую определённый интеграл методом Монте-Карло, чтобы с вероятностью  $p = 0,9$  обеспечить точность вычисления интеграла  $\varepsilon = 0,01$ .

4. Воспользовавшись программой, провести 10 серий вычислений, после чего найти среднее арифметическое.

5. Сделать выводы по работе.

*При сдаче лабораторной работы демонстрируется:*

1. Запускающийся exe-файл разработанной программы + код программы;

2. График выбранной функции с отмеченными пределами интегрирования;
3. Качественный отчёт.

### ***Пример выполнения***

Вычислим интеграл Пуассона  $\int_a^b e^{-x^2} dx$ . Как известно, функция  $y = e^{-x^2}$  не имеет первообразной в элементарных функциях, т. е. интеграл нельзя вычислить непосредственно.

График функции  $y = e^{-x^2}$  представлен на рис. 4.2.

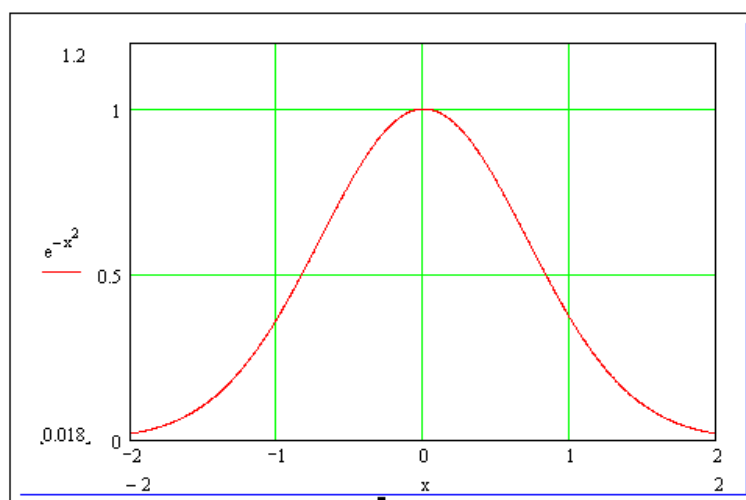


Рис. 4.2

Выберем пределы интегрирования: пусть  $a = -1$ ;  $b = 2$ . Т. е. необходимо найти площадь закрашенной фигуры (рис. 4.3).

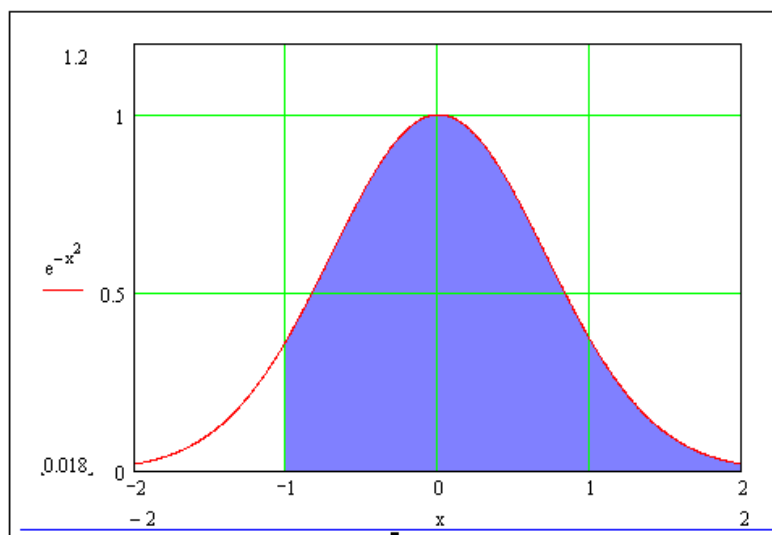


Рис. 4.3

Для этого будем генерировать (программа) пары чисел  $(x_i, y_i)$  ( $1 \leq i \leq n$ ) в область  $D$ , указанную на рис. 4.4.

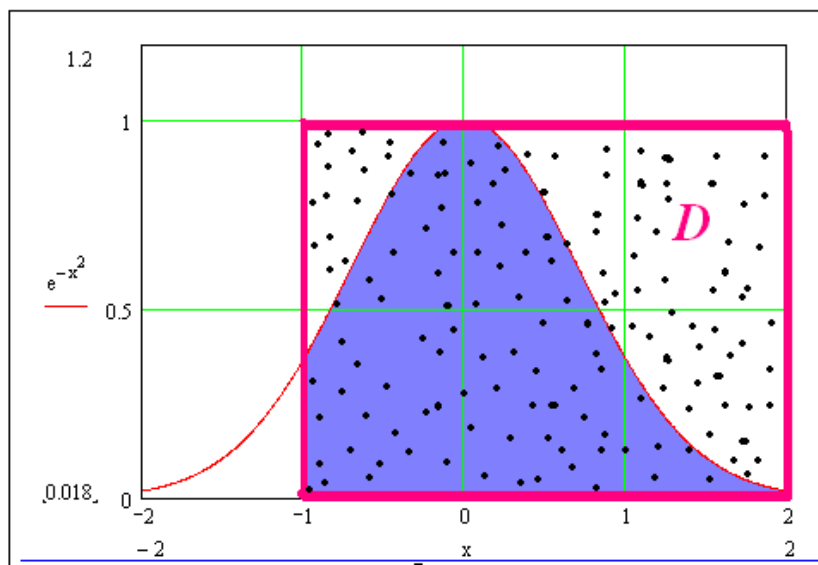


Рис. 4.4

Мера множества  $\mu(D)$  равна 3. Так как нижний предел интегрирования функции  $a = -1$ , а верхний  $b = 2$ , следовательно,  $x_i$  – случайные числа, которые генерируются на отрезке  $[-1; 2]$ . Так как наибольшее значение, которое принимает функция  $y = e^{-x^2}$  на отрезке  $[-1; 2]$ , есть  $y(0) = 1$ , то  $y_i$  – случайные числа, которые генерируются на отрезке  $[0; 1]$ .

Условие попадания точки под график функции:  $y_i \leq f(x_i)$ .

Тогда  $\int_{-1}^2 e^{-x^2} dx \approx \frac{m}{n} \cdot \mu(D)$ , где  $m$  – количество точек, попавших под график функции;  $n$  – количество всех точек;  $\mu(D)$  – мера множества  $D$ .

Количество всех точек  $n$  необходимо найти из условия, что с вероятностью  $p = 0,9$  точность вычисления интеграла должна быть не менее  $\varepsilon = 0,01$  (см. пункт б) первого этапа). Таким образом, в результате выполнения всех пунктов задания был получен результат  $\int_{-1}^2 e^{-x^2} dx \approx 1,62815$ . Истинное же значение данного интеграла 1,62891.

При выполнении лабораторной работы подынтегральная функция, соответствующая условиям варианта (см. прил. 2), выбирается студентами произвольно. Например, если в задании требуется выбрать функцию, в которой



присутствуют натуральный логарифм и  $x^x$ , то можно выбрать функцию вида  $y = \operatorname{arcsinh}(x-1) \cdot \frac{\sqrt{\ln(5x+1)}}{x^x}$ , которая, очевидно, удовлетворяет требованиям.

Пределы интегрирования также выбираются произвольно, например, от 1 до 3 (рис. 4.5).

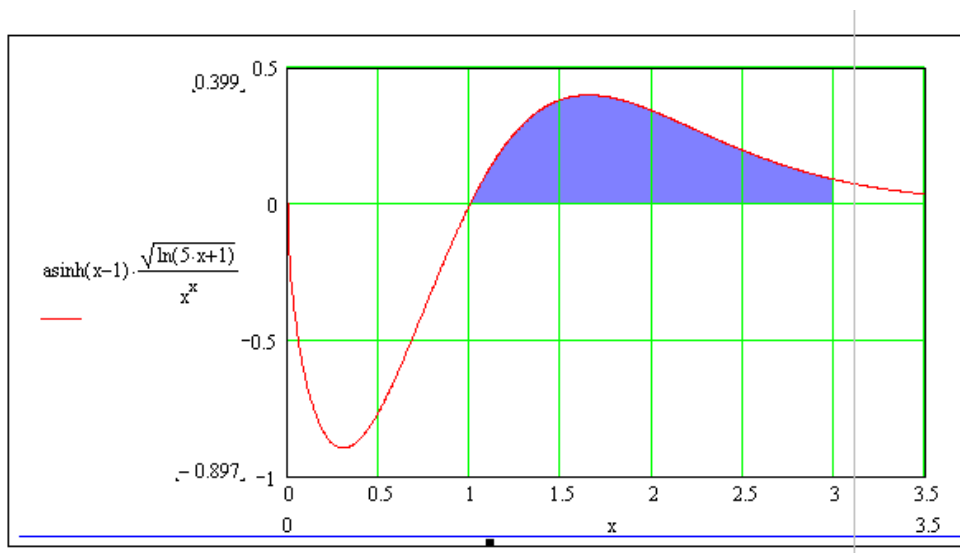


Рис. 4.5

## Занятие 5

### Применение Excel и системы STATISTICA при решении задач по теории вероятностей

При решении комбинаторных задач и задач на вычисление вероятности подчас требуется выполнение большого количества однообразных рутинных вычислений. Рассмотрим возможности применения Excel и системы STATISTICA в качестве средства вычислений.

#### Цели занятия

**Знать** – простейшие модели и формулы комбинаторики и теории вероятностей.

**Уметь** – используя средства Excel и STATISTICA, решать задачи на вычисление вероятности и количества комбинаций.

**Владеть** – навыками использования Excel и STATISTICA при вычислении значений вероятности.

#### 1. Применение Excel при решении комбинаторных задач

**Комбинаторика** изучает количества комбинаций, подчиненных определенным условиям, которые можно составить из элементов, безразлично какой природы, заданного конечного множества.

**Перестановками** называют комбинации, состоящие из одних и тех же  $n$  различных элементов и отличающиеся только порядком их расположения. Число всех возможных перестановок  $P_n = n!$ .

**Размещением** называют комбинации, составленные из  $n$  различных элементов по  $m$  элементов, которые отличаются либо составом элементов, либо их порядком. Число всех возможных размещений  $A_n^m = \frac{n!}{(n-m)!}$ .

**Сочетанием** называют комбинации, составленные из  $n$  различных элементов по  $m$  элементов, которые отличаются хотя бы одним элементом.

Число сочетаний  $C_n^m = \frac{n!}{m!(n-m)!}$ .

**Пример 1.** Порядок выступления 10 участников соревнований определяется жребием. Сколько вариантов при этом возможно?

**Решение.** Каждый вариант отличается только порядком, поэтому является перестановкой из 10 элементов ( $n=10$ ). Тогда количество вариантов равно  $P_{10} = 10! = 3628800$ . Для вычисления в Excel используем функцию ПЕРЕСТ с равными значениями аргументов. Формула =ПЕРЕСТ(10; 10) дает значение 3628800.

**Пример 2.** Сколькими способами можно выбрать старосту и профорга в группе из 15 студентов?

**Решение.** Комбинации из 2-х элементов различаются составом и порядком, т.е. являются размещениями с  $n = 15$  и  $m = 2$ . Тогда количество вариантов определяется по формуле  $A_{15}^2 = \frac{15!}{(15-2)!} = 210$ . Для вычисления в Excel используем формулу =ПЕРЕСТ(15; 2), которая дает значение 210.

**Пример 3.** Сколькими способами можно выбрать двух делегатов на студенческую конференцию в группе из 15 студентов?

**Решение.** В данном случае порядок элементов не важен, наборы различаются только составом, т.е. являются сочетаниями с параметрами  $n=15$  и  $m=2$ . Тогда количество вариантов определяется по формуле  $C_{15}^2 = \frac{15!}{2!(15-2)!} = \frac{A_{15}^2}{P_2} = 105$ . Для вычисления в Excel используем формулу =ПЕРЕСТ(15; 2)/ПЕРЕСТ(2; 2), которая дает значение 105.

## 2. Решение задач на вычисление вероятности в Excel

**Пример 4.** В партии 20 изделий, из них 5 бракованных. Найти вероятность того, что в выборке из 4 изделий ровно одно бракованное.

**Решение.** В данной задаче, прежде всего, определим значения параметров: *число\_успехов\_в\_выборке* = 1; *размер\_выборки* = 4; *число\_успехов\_в\_совокупности* = 5; *размер\_совокупности* = 20.

Искомую вероятность можно рассчитать с помощью функции =ГИПЕРГЕОМЕТ(1; 4; 5; 20), которая дает значение 0,4696.

Если производится несколько испытаний, причем вероятность события  $A$  в каждом испытании не зависит от исходов других испытаний, то такие испытания называют **независимыми относительно события  $A$** .

Пусть производится  $n$  независимых испытаний, в каждом из которых событие  $A$  может появиться либо не появиться. Вероятность события  $A$  в каждом испытании одна и та же, а именно равна  $p$ . Следовательно, вероятность ненаступления события  $A$  в каждом испытании также постоянна и равна  $q = 1 - p$ .

Вероятность того, что при  $n$  повторных независимых испытаниях событие  $A$  осуществится ровно  $k$  раз вычисляется по **формуле Бернулли**:  $P_n(k) = C_n^k p^k q^{n-k}$ .

Для нахождения наиболее вероятного числа успехов  $k_0$  по заданным  $n$  и  $p$  можно воспользоваться неравенствами  $np - q \leq k_0 \leq np + p$  или правилом: если число  $np + p$  не целое, то  $k_0$  равно целой части этого числа.

В случае, если  $n$  велико,  $p$  мало, а  $\lambda = np \leq 10$ , используют **асимптотическую формулу Пуассона** вычисления вероятности наступления события  $A$  ровно  $k$  раз при  $n$  повторных независимых испытаниях:  $P_n(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ .

**Пример 5.** Вероятность того, что расход электроэнергии на протяжении одних суток не превысит установленной нормы, равна  $p = 0,75$ . Найти вероятность того, что в ближайшие 6 суток расход электроэнергии в течение 4 суток не превысит нормы.

**Решение.** Вероятность нормального расхода электроэнергии на протяжении каждых из 6 суток постоянна и равна  $p = 0,75$ . Следовательно, вероятность перерасхода электроэнергии в каждые сутки также постоянна и равна  $q = 1 - p = 1 - 0,75 = 0,25$ . Искомая вероятность по формуле Бернулли равна  $P_6(4) = C_6^4 p^4 q^{6-4} = 0,297$ . Для вычисления в Excel используем формулу

=**БИНОМРАСП(4; 6; 0,75; 0)**, которая дает значение 0,297. При этом определены следующие значения параметров: **число\_успехов** = 4; **число\_испытаний** = 6; **вероятность\_успеха** = 0,75; **интегральная** = 0. Подробно с синтаксисом функции **БИНОМРАСП** можно ознакомиться с помощью справки.

**Пример 6.** Телефонная станция обслуживает 400 абонентов. Для каждого абонента вероятность того, что в течение часа он позвонит на станцию, равна 0,01. Найти вероятность, что в течение часа ровно 5 абонентов позвонят на станцию.

**Решение.** Так как  $p = 0,01$  мало и  $n = 400$  велико, то будем пользоваться приближенной формулой Пуассона при  $\lambda = 400 \cdot 0,01 = 4$ . Тогда  $P_{400}(5) \approx \frac{4^5}{5!} e^{-4} \approx 0,156293$ . Для вычисления в Excel используем формулу **=ПУАССОН(5; 4; 0)**, которая дает значение 0,156293. При этом определены следующие значения параметров: **количество\_событий** = 5; **среднее** ( $\lambda$ ) = 4; **интегральная** = 0. Подробно с синтаксисом функции **ПУАССОН** можно ознакомиться в справке.

В случае, когда число повторных испытаний большое и формула Бернулли неприменима, используют формулы Лапласа.

**Локальная теорема Лапласа.** Если вероятность  $p$  появления события  $A$  в каждом испытании постоянна и отлична от нуля и единицы, то вероятность  $P_n(k)$  того, что событие  $A$  появится в  $n$  испытаниях ровно  $k$  раз, приближенно равна (тем точнее, чем больше  $n$ ) значению функции  $P_k(n) = \frac{1}{\sqrt{npq}} \varphi(x)$ , где

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x = \frac{k - np}{\sqrt{npq}}.$$

Имеются таблицы, в которых помещены значения функции  $\varphi(x)$ .

**Интегральная теорема Лапласа.** Если вероятность  $p$  наступления события  $A$  в каждом испытании постоянна и отлична от нуля и единицы, то вероятность  $P_n(k_1, k_2)$  того, что событие  $A$  появится в  $n$  испытаниях от  $k_1$  до  $k_2$  раз, приближенно равна определенному интегралу:

$$P_n(k_1, k_2) = \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-\frac{z^2}{2}} dz, \quad \text{где } x' = \frac{k_1 - np}{\sqrt{npq}}, \quad x'' = \frac{k_2 - np}{\sqrt{npq}}.$$

При решении задач, требующих применения интегральной теоремы Лапласа, пользуются специальными таблицами для интеграла  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$ , тогда  $P_n(k_1, k_2) = \Phi(x'') - \Phi(x')$ .

**Пример 7.** Найти вероятность того, что событие  $A$  наступит ровно 80 раз в 400 испытаниях, если вероятность появления этого события в каждом испытании равна 0,2.

**Решение.** По условию  $n = 400$ ;  $k = 80$ ;  $p = 0,2$ ;  $q = 0,8$ . Воспользуемся асимптотической формулой Лапласа:  $x = \frac{80 - 400 \cdot 0,2}{\sqrt{400 \cdot 0,2 \cdot 0,8}} = 0$ ,  $\varphi(0) = 0,3989$ ,

$P_{80}(400) = \frac{1}{\sqrt{400 \cdot 0,2 \cdot 0,8}} 0,3989 = 0,04986$ . Для вычисления в Excel используем формулу **=НОРМРАСП(80; 80; 8; 0)**, которая дает значение 0,04986. При этом определены следующие значения параметров:  $k = 80$ ; **среднее**  $= np = 80$ ; **стандартное\_откл**  $= \sqrt{npq} = \sqrt{400 \cdot 0,2 \cdot 0,8} = 8$ , **интегральная**  $= 0$ . Подробно с синтаксисом функции **НОРМРАСП** можно ознакомиться с помощью справки.

**Пример 8.** Вероятность того, что деталь не прошла проверку ОТК, равна 0,2. Найти вероятность того, что среди 400 случайно отобранных деталей окажется непроверенных от 70 до 100 деталей.

**Решение.** Воспользуемся интегральной формулой Лапласа:  $n = 400$ ;  $k_1 = 70$ ;  $k_2 = 100$ ;  $p = 0,2$ ;  $q = 0,8$ ;  $x' = \frac{70 - 80}{\sqrt{400 \cdot 0,2 \cdot 0,8}} = -1,25$ ,  $x'' = \frac{100 - 80}{\sqrt{400 \cdot 0,2 \cdot 0,8}} = 2,5$ . Так

как функция  $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$  является нечетной, то  $P_{400}(70; 100) = \Phi(2,5) + \Phi(1,25) = 0,4938 + 0,3944 = 0,8882$ .

Для вычисления в Excel используем формулу нормального распределения **=НОРМРАСП(100; 80; 8; 1) – НОРМРАСП(70; 80; 8; 1)**, которая дает значение 0,8882. При этом параметр **интегральная**  $= 1$ , остальные значения параметров определяются аналогично примеру, рассмотренному выше.

### 3. Решение задач на вычисление вероятности в системе STATISTICA

Примеры решения задач и задания для самостоятельной работы приведены в [2] и на сайте [www.statistica.ru](http://www.statistica.ru).

#### Задание для самостоятельной работы

1. Решить задачи, сделать проверку, выполняя вычисления в Excel или STATISTICA.

**Задача 1.** В группе 12 студентов, среди которых 3 отличника. По списку наудачу отобраны 9 студентов. Найдите вероятность того, что среди отобранных студентов 2 отличника.

**Задача 2.** В новом микрорайоне поставлено 1000 кодовых замков на входных дверях домов. Вероятность выхода из строя одного замка в течение месяца равна 0,001. Найдите вероятность того, что за месяц откажут:

а) 2 замка; б) не более 3-х замков; в) не менее 2-х замков; г) хотя бы один замок.

**Задача 3.** Сто станков работают независимо друг от друга, причем вероятность бесперебойной работы каждого из них в течение смены равна 0,8. Найдите вероятность того, что в течение смены бесперебойно проработают:

а) 85 станков; б) от 75 до 85 станков.

2. На занятии 3 была рассмотрена статистическая функция **Мастера функций Excel НОРМОБР**. Какие задачи теории вероятностей позволяет решать функция **НОРМОБР**?

3. Используя систему справки и литературу по анализу данных в Excel, изучите синтаксис функций **ПЕРЕСТ**, **БИНОМРАСП**, **ГИПЕРГЕОМЕТ**, **ПУАССОН**, **НОРМРАСП**, **НОРМОБР**.

### **Задание для самостоятельного изучения**

#### **Вероятностный калькулятор в системе STATISTICA**

Вероятностный калькулятор в системе STATISTICA – это средство, позволяющее максимально быстро построить графики наиболее употребляемых функций распределения и их плотностей, вычислить процентные точки. Применение вероятностного калькулятора заменяет использование таблиц распределений. Приемы работы с вероятностным калькулятором приведены в книге В. Боровикова [2].

#### **Цели задания**

**Знать** – способы задания, свойства и характеристики важнейших непрерывных распределений.

**Уметь** – производить вычисления, используя возможности средства «Вероятностный калькулятор».

**Владеть** – навыками применения вероятностного калькулятора при решении задач теории вероятностей и математической статистики.

Запустите модуль **Basic statistics/Tables** (Основные статистики/таблицы) из Переключателя модулей. В стартовой панели данного модуля выделите курсором строку **Probability calculator** (Вероятностный калькулятор) и нажмите кнопку **ОК** (рис. 5.1).

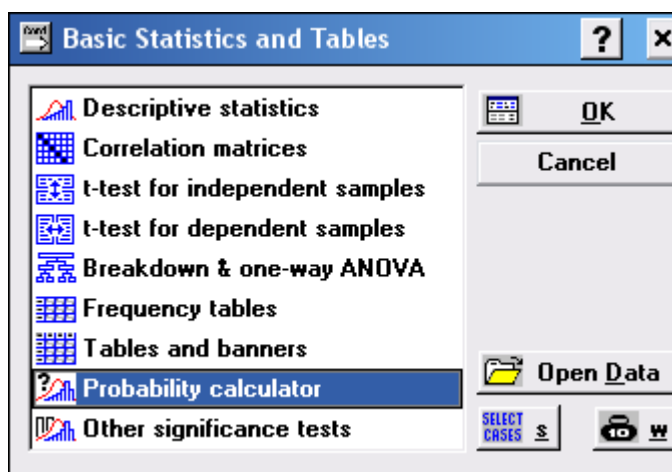


Рис. 5.1

Откроется окно **Probability Distribution Calculator** (Калькулятор вероятностных распределений), в левой части которого приводится список распределений (рис. 5.2). Наиболее важными для анализа данных являются нормальное распределение, хи-квадрат распределение, t-распределение Стьюдента и F-распределение Фишера.

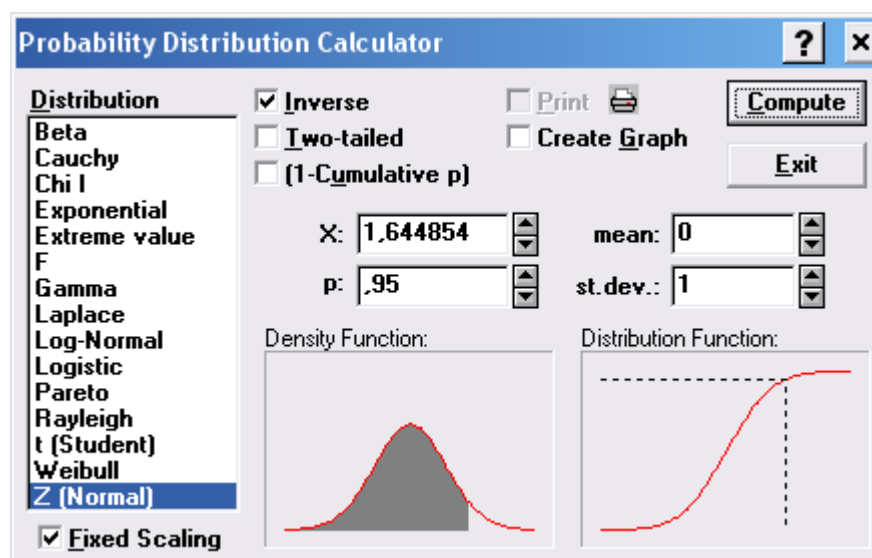


Рис. 5.2

## 1. Изучение свойств нормального распределения

Нормальное распределение имеет два параметра:  $\mu$  – математическое ожидание (**mean**) и  $\sigma$  – среднее квадратическое отклонение (**st. dev.**). Нормальное распределение с  $\mu = 0$ ,  $\sigma = 1$  называется стандартным нормальным

распределением. В списке вероятностного калькулятора выделите поле **Z (Normal)**.

**Упражнение 1.** а) Изменяя значения математического ожидания (поле **mean**), проанализируйте влияние данного параметра на положение нормальной кривой.

б) При фиксированном значении математического ожидания изменяйте значение  $\sigma$  (поле **st. dev.**), например  $\sigma = 0,5$ ,  $\sigma = 1$ ,  $\sigma = 2$ . Сделайте вывод о том, как изменяется форма графика плотности распределения.

## 2. Изучение свойств распределения хи-квадрат

Случайная величина, имеющая распределение хи-квадрат, определяется как сумма квадратов  $k$  независимых нормальных случайных величин. При этом число  $k$  называется числом степеней свободы и является единственным параметром данного распределения. Распределение хи-квадрат сосредоточено на положительной полуоси. В списке вероятностного калькулятора выделите **Chi I** – распределение хи-квадрат.

**Упражнение 2.** а) Изменяя значение параметра  $k$  (поле **df**), например  $k = 2$ ,  $k = 7$ ,  $k = 25$ , проанализируйте влияние данного параметра на форму графика плотности.

б) При фиксированном значении  $k = 7$  задайте значение поля **p**, равное 0,95. Сравните значение поля **Chi I** с табличным значением 0,95-квантили для хи-квадрат распределения с 7 степенями свободы.

## 3. Изучение свойств t-распределения Стьюдента

Данное распределение имеет один параметр  $k$  (число степеней свободы) и сосредоточено на всей действительной оси симметрично относительно 0. То есть математическое ожидание равно 0, а дисперсия равна  $k/(k - 2)$ . В списке вероятностного калькулятора выделите поле **t (Student)**.

**Упражнение 3.** а) Постройте график плотности распределения Стьюдента с 5 степенями свободы (поле **df**). По уровню **p=0,95** найдите значение поля **t** и сравните его с табличным значением 0,95-квантили для распределения Стьюдента с 5 степенями свободы.

б) Постройте график плотности t-распределения Стьюдента с 30 степенями свободы и сравните его с графиком стандартного нормального распределения. Сделайте вывод.



#### 4. Изучение свойств F-распределения Фишера

Данное распределение имеет два параметра  $m$  и  $n$  – количество степеней свободы. Случайная величина, имеющая F-распределение с парой степеней свободы  $m$  и  $n$ , определяется как отношение двух независимых случайных величин, имеющих распределение хи-квадрат со степенями свободы  $m$  и  $n$  с умножением на нормировочный множитель  $n/m$ . F-распределение сосредоточено на положительной полуоси. В списке вероятностного калькулятора выделите поле **F** – распределение Фишера.

**Упражнение 4.** Постройте график плотности F-распределения Фишера со степенями свободы 5 и 10 (поля **df1** и **df2**). По уровню  $p=0,95$  найдите значение поля **F** и сравните его с табличным значением 0,95-квантили для распределения Фишера соответствующими значениями степеней свободы.

**Упражнение 5.** Изучив формулы плотности, рассмотрите вид и свойства графиков для других распределений, представленных в вероятностном калькуляторе. Например, логнормальное распределение, распределение Коши, распределение Вейбулла, распределение Парето, экспоненциальное распределение.

#### Занятие 6

##### Законы больших чисел. Теорема Чебышева. Центральная предельная теорема Ляпунова. Лабораторная работа № 3

Данная тема важна для понимания методов математической статистики. Она включает ряд утверждений, устанавливающих при определенных условиях устойчивость частности и средней арифметической (теоремы Бернулли, Пуассона, Чебышева, Маркова и др.) или устойчивость закона распределения (теорема Ляпунова). При изучении каждого вопроса темы важно уяснить условия применения.

##### Цели занятия

- *Формирование навыков стохастического моделирования.*
- *Уяснение сущности законов больших чисел, их важности для статистического вывода.*
- *Экспериментальная проверка влияния условий применимости и последствий их нарушения.*

**Теорема Чебышева.** Если независимые случайные величины  $X_1, \dots, X_n$  имеют одинаковые математические ожидания, равные  $a$ , и их дисперсии ограничены одной и той же постоянной  $C$ , то

$$P\left(\left|\frac{X_1 + \dots + X_n}{n} - a\right| \leq \varepsilon\right) \geq 1 - \frac{C}{n\varepsilon^2} \quad \text{или} \quad \lim_{n \rightarrow \infty} P\left(\left|\frac{X_1 + \dots + X_n}{n} - a\right| \leq \varepsilon\right) = 1,$$

то есть, при увеличении количества слагаемых  $n$  средняя арифметическая случайных величин  $\bar{X} = (X_1 + \dots + X_n)/n$  сходится по вероятности к математическому ожиданию  $a$ .

Выполнение закона больших чисел отражает предельную устойчивость средних арифметических случайных величин: при большом числе испытаний они практически перестают быть случайными и совпадают со своими средними значениями.

**Теорема Ляпунова.** Пусть  $X_1, \dots, X_n$  – независимы и одинаково распределены,  $E(X_i) = a$  и  $V(X_i) = \sigma^2$  ( $i = 1, \dots, n$ ). Пусть  $Y = X_1 + \dots + X_n$ . Тогда при большом значении количества слагаемых  $n$  функция плотности случайной величины  $Y$  сойдется к нормальному закону распределения.

### **Лабораторная работа № 3. Экспериментальное изучение законов больших чисел**

**Цель работы:** экспериментальная проверка выполнения теоремы Чебышева и центральной предельной теоремы Ляпунова.

Выполнение лабораторной работы разобьем на два этапа.

*На первом этапе* для проверки теоремы Чебышева требуется:

1) сгенерировать в Excel или Statistica 10 столбцов по 100 чисел, имеющих заданное распределение: нормальное  $N(a, \sigma)$ , экспоненциальное  $E(\lambda)$  или равномерное на  $[a, b]$ ;

2) вычислить средние арифметические  $\bar{X}$  и убедиться, что их значения близки к математическим ожиданиям сгенерированных случайных величин.

Распределение Коши, как известно, не имеет ограниченной дисперсии: возможны редкие выбросы большой величины. Условия теоремы Чебышева не выполняются, и  $\bar{X}$  не стремится с ростом  $n$  к какой-либо константе. Для проверки этого факта следует сгенерировать 10 столбцов по 100 чисел, имеющих распределение Коши с функцией плотности  $f(x) = 1/(\pi(1 + x^2))$ , вычислить средние арифметические  $\bar{X}$  и убедиться в отсутствии сходимости этих величин.

На втором этапе для проверки справедливости центральной предельной теоремы Ляпунова методом графического анализа требуется:

1) сгенерировать в Excel или Statistica 8 столбцов по 100 случайных чисел  $x_1, \dots, x_8$ , имеющих заданное распределение: нормальное  $N(a, \sigma)$ , экспоненциальное  $E(\lambda)$  или равномерное на  $[a, b]$ ;

2) вычислить  $y_1 = x_1 + x_2, y_2 = y_1 + x_3, \dots, y_7 = y_6 + x_8$ ;

3) построив гистограммы для сумм  $y_1, \dots, y_7$ , проверить, что при увеличении количества слагаемых в сумме, ее распределение приближается к нормальному. Какова при этом роль вида распределения слагаемых  $x_1, \dots, x_8$ ?

Сделать выводы по работе.

## Занятие 7

### Проверка гипотез. Критерий хи-квадрат проверки гипотез

Статистическое исследование направлено на поиск закономерностей, описывающих случайные явления. Исследования начинаются с разведочного анализа: получения выборки, ее упорядочения, группировки, вычисления числовых характеристик, графического представления. На основе полученных результатов можно проводить более глубокие статистические исследования, к которым относится проверка статистических гипотез.

### Цели занятия

**Знать** – определения понятий, связанных с проверкой статистических гипотез, общую схему проверки статистических гипотез, примеры ситуаций и задач, возникающих в ходе анализа данных, решение которых требует применения критерия хи-квадрат, алгоритмы проверки статистических гипотез с применением критерия хи-квадрат.

**Уметь** – осуществлять проверку гипотез, делать выводы.

**Владеть** – навыками проверки статистических гипотез с использованием системы STATISTICA.

**Статистической** называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

**Нулевой (основной)** называют выдвинутую (проверяемую) гипотезу.

**Конкурирующей (альтернативной)** называют гипотезу, которая противоречит нулевой.

**Простая гипотеза** – та гипотеза, в которой на проверку выдвигается только один параметр. **Сложная гипотеза**, в которой на проверку выдвигаются два и более параметра.

**Ошибка первого рода** состоит в том, что будет отвергнута правильная гипотеза. **Ошибка второго рода** состоит в том, что будет принята неправильная гипотеза.

**Статистическим критерием** называют случайную величину  $K$ , которая будет служить для проверки нулевой гипотезы.

Наблюдаемым значением  $K_{набл}$  называют значение критерия, вычисленное по выборке.

**Критической областью** называют совокупность значений критерия, при котором нулевую гипотезу отвергают. **Область принятия гипотезы** – совокупность значений критерия, при котором гипотезу принимают.

Основной принцип проверки статистических гипотез: если наблюдаемое значение критерия принадлежит критической области, гипотезу отвергают. Если наблюдаемое значение критерия принадлежит области принятия гипотезы, то ее принимают.

**Критическими точками (границами)** называют точки, отделяющие критическую область от области принятия гипотезы.

Области бывают односторонние (право- или лево-) и двусторонние.

Правосторонняя – это критическая область, определяемая неравенством  $K > K_{кр}$ . Левосторонняя:  $K < K_{кр}$ . Двусторонняя:  $K < K_1, K > K_2$ .

Критерий хи-квадрат Пирсона является весьма общим методом построения тестов для проверки различных гипотез. Рассмотрим исходную схему.

## 1. Проверка простой гипотезы о вероятностях

Обозначим:

-  $A_1, \dots, A_m$  –  $m$  возможных исходов некоторого опыта;

-  $p_1, \dots, p_m$  – вероятности соответствующих исходов,  $\sum_{i=1}^m p_i = 1$ ;

-  $n$  – число независимых повторений опыта;

-  $v_1, \dots, v_m$  – число появлений соответствующих исходов в  $n$  опытах,

где  $\sum_{i=1}^m v_i = n$ ;

-  $p_1^0, \dots, p_m^0$  – гипотетические значения вероятностей,  $p_i^0 > 0$  и  $\sum_{i=1}^m p_i^0 = 1$ .

Требуется по наблюдениям  $\nu_1, \dots, \nu_m$  проверить гипотезу **H** о том, что вероятности  $p_1, \dots, p_m$  имеют значения  $p_1^0, \dots, p_m^0$ , т.е.

$$\mathbf{H}: p_i = p_i^0 \quad (i=1, \dots, m).$$

Оценками для  $p_1, \dots, p_m$  являются  $\hat{p}_1 = \nu_1/n, \dots, \hat{p}_m = \nu_m/n$ . Мерой расхождения между гипотетическими вероятностями  $p_1^0, \dots, p_m^0$  и эмпирическими вероятностями  $\hat{p}_1, \dots, \hat{p}_m$  принимается величина

$$X^2 = n \sum_{i=1}^m p_i^0 \left( \frac{\hat{p}_i - p_i^0}{p_i^0} \right)^2,$$

которая с точностью до множителя  $n$  есть усредненное с весами  $p_i^0$  значение квадрата относительного отклонения значений  $\hat{p}_i$  от  $p_i^0$ . Статистика  $X^2$  называется статистикой хи-квадрат Пирсона. Для ее вычисления используются две формулы:

$$X^2 = \sum_{i=1}^m \frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^m \frac{\nu_i^2}{np_i^0} - n. \quad (7.1)$$

Условно статистику можно записать так:

$$X^2 = \sum \frac{(H - T)^2}{T},$$

где  $H$  – наблюдаемые частоты  $\nu_i$ ;  $T$  – теоретические (ожидаемые) частоты  $np_i^0$ .

Поскольку по закону больших чисел  $\hat{p}_i \rightarrow p_i$  при  $n \rightarrow \infty$  то

$$\sum_{i=1}^m p_i^0 \left( \frac{\hat{p}_i - p_i^0}{p_i^0} \right)^2 \rightarrow \sum_{i=1}^m \frac{(p_i - p_i^0)^2}{p_i^0}.$$

Последняя величина равна 0, если верна **H**; если же **H** не верна, то  $X^2 \rightarrow \infty$

Процедура проверки гипотезы состоит в том, что если величина  $X^2$  приняла «слишком большое» значение, т.е. если

$$X^2 \geq h, \quad (7.2)$$

то гипотеза **H** отклоняется; если это не так, будем говорить, что наблюдения не противоречат гипотезе. На вопрос, что означает «слишком большое» значение, отвечает теорема К. Пирсона.

**Теорема К. Пирсона.** Если гипотеза **H** верна и  $p_i^0 \geq 0$  ( $i=1, \dots, m$ ), то при  $n \rightarrow \infty$  распределение статистики  $X^2$  асимптотически подчиняется распределению хи-квадрат с  $m-1$  степенями свободы, т.е.

$$P\{X^2 < x / \mathbf{H}\} \rightarrow F_{m-1}(x) \equiv P\{\chi_{m-1}^2 < x\}.$$

Порог  $h$  выберем из условия: вероятность ошибки первого рода должна быть малой – равной выбираемому значению уровня значимости  $\alpha$ :

$$P\{\text{отклонить } \mathbf{H} / \mathbf{H} \text{ верна}\} = P\{X^2 \geq h / \mathbf{H}\} \equiv P\{\chi_{m-1}^2 \geq h\} = \alpha,$$

откуда

$$h = Q(1 - \alpha, m - 1). \quad (7.3)$$

$h$  – квантиль уровня  $1 - \alpha$  распределения хи-квадрат с  $m - 1$  степенями свободы.

Процедура (7.2) – (7.3) проверки **H** может быть записана иначе: гипотеза **H** отклоняется, если

$$P\{\chi_{m-1}^2 \geq X^2\} \leq \alpha, \quad (7.4)$$

т.е. если мала вероятность получения (при справедливости **H**) такого же расхождения, как в опыте (т.е.  $X^2$ ), или ещё большего. Вероятность слева в (7.4) называется минимальным уровнем значимости (при любом значении  $\alpha$ , большем  $P\{\chi_{m-1}^2 \geq X^2\}$ , гипотеза, очевидно, отклоняется).

**Замечание.** Теорему Пирсона можно применять, если все ожидаемые частоты

$$np_i^0 \geq 10, \quad i = 1, \dots, m. \quad (7.5a)$$

Если  $m$  порядка десяти и более, достаточно выполнения

$$np_i^0 \geq 4, \quad i = 1, \dots, m. \quad (7.5b)$$

Если (7.5) не выполняется, необходимо некоторые исходы  $A_i$  объединять.

## 2. Проверка сложной гипотезы о вероятностях

Пусть  $A_1, \dots, A_m$  –  $m$  исходов некоторого опыта,  $n$  – число независимых повторений опыта,  $\nu_1, \dots, \nu_m$  – количества появлений исходов. Проверяемая гипотеза **H** предполагает, что вероятности исходов  $P(A_i)$  являются известными функциями  $p_i(a)$   $k$ -мерного параметра  $a = (a_1, \dots, a_k)$ , т.е.

$$\mathbf{H}: P(A_i) = p_i(a), \quad (i=1, \dots, m),$$

но значение  $a$  неизвестно.

Для проверки гипотезы **H** определим статистику

$$\tilde{X}^2 = \min_a \sum_{i=1}^m \frac{(v_i - np_i(a))^2}{np_i(a)}. \quad (7.6)$$

По теореме Фишера, если **H** верна, то при  $n \rightarrow \infty$  распределение статистики  $\tilde{X}^2$  асимптотически подчиняется распределению хи-квадрат с числом степеней свободы  $f = m - 1 - k$ , и потому **отклоняем H**, если

$$\tilde{X}^2 \geq h, \quad (7.7)$$

где  $h = Q(1 - \alpha, f)$  – квантиль уровня  $1 - \alpha$  распределения хи-квадрат с числом степеней свободы  $f$ ; такой порог обеспечивает выбранный уровень  $\alpha$  вероятности  $P\{\text{отклонить H} / \text{H верна}\}$  ошибки 1-го рода. Если (7.7) не выполняется, делаем вывод: **наблюдения не противоречат гипотезе**. Распределению хи-квадрат с  $f$  степенями свободы асимптотически подчиняется также статистика

$$\tilde{X}^2 = \sum_{i=1}^m \frac{(v_i - np_i(\hat{a}))^2}{np_i(\hat{a})}, \quad (7.8)$$

где  $\hat{a}$  – оценка максимального правдоподобия для  $a$ , и потому в (7.7) может быть использована статистика (7.8) вместо (7.6). Процедуру (7.7) можно записать иначе: если

$$P\{\chi_f^2 \geq X^2\} \leq \alpha, \quad (7.9)$$

то гипотеза **H** отклоняется.

### 3. Проверка гипотезы о типе распределения

Пусть требуется проверить гипотезу о том, что выборка  $x_1, \dots, x_n$  извлечена из совокупности, распределенной по некоторому закону, известному с точностью до  $k$ -мерного параметра  $a = (a_1, \dots, a_k)$ . Оказываются теоретически обоснованными следующие действия: разобьем весь диапазон наблюдений на  $m$  интервалов, определим значения  $v_i$  (число наблюдений в  $i$ -м интервале), получим значение оценки  $\hat{a}$  минимизацией (7.6) или методом максимального правдоподобия, определим вероятности  $p_i(\hat{a})$  попадания в  $i$ -й интервал, вычислим (7.6) или (7.8) и примем решение по (7.7).

**Пример 7.1. Проверка нормальности распределения.** Проверим гипотезу о нормальном законе распределения размеров головок заклепок, сделанных на одном станке, по выборке объема  $n = 200$ ; измерения приведены в табл. 1. Оценками для  $a$  (среднего) и  $\sigma$  (стандартного отклонения) являются

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{и} \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

### Выполнение в пакете STATISTICA

Результаты измерения диаметров заклепок занесем в таблицу с одним столбцом ( $d$ ) и 200 строками; соответствующий файл назовем, например, *Diamz.sta*. Работаем в модуле *Nonparametric Statistics* (непараметрическая статистика), *Distribution Fitting* (подбор распределения). В поле *Continuous Distributions: Normal; Variable: d*; в поле *Plot distribution: Frequency distribution* (частоты распределения) и отказываемся от теста Колмогорова - Смирнова: **OK**. Наблюдаем оценки параметров **Mean: 13,42**, **Variance: 0,018**, соглашаемся с параметрами группирования (в частности, с числом групп **Number of categories: 19**) – **OK**.

**Таблица 7.1.** Диаметры 200 головок заклепок, мм

13.39	13.33	13.56	13.38	13.43	13.37	13.53	13.40	13.25	13.37
13.28	13.34	13.50	13.38	13.38	13.45	13.47	13.62	13.45	13.39
13.53	13.58	13.32	13.27	13.42	13.40	13.57	13.46	13.33	13.40
13.57	13.36	13.43	13.38	13.26	13.52	13.35	13.29	13.48	13.43
13.40	13.39	13.50	13.52	13.39	13.39	13.46	13.29	13.55	13.31
13.29	13.33	13.38	13.61	13.55	13.40	13.20	13.31	13.46	13.13
13.43	13.51	13.50	13.38	13.44	13.62	13.42	13.54	13.31	13.58
13.41	13.49	13.42	13.45	13.34	13.47	13.48	13.59	13.20	14.56
13.55	13.44	13.50	13.40	13.48	13.29	13.31	13.42	13.32	13.48
13.43	13.26	13.58	13.38	13.48	13.45	13.29	13.32	13.24	13.38
13.34	13.14	13.31	13.51	13.59	13.32	13.52	13.57	13.62	13.29
13.23	13.37	13.64	13.30	13.40	13.58	13.24	13.32	13.52	13.50
13.43	13.58	13.63	13.48	13.34	13.37	13.18	13.50	13.45	13.60
13.38	13.33	13.57	13.28	13.32	13.40	13.40	13.33	13.20	13.44
13.34	13.54	13.40	13.47	13.28	13.41	13.39	13.48	13.42	13.46
13.28	13.46	13.37	13.53	13.43	13.30	13.45	13.40	13.45	13.40
13.33	13.39	13.56	13.46	13.26	13.35	13.42	13.36	13.44	13.41
13.43	13.51	13.51	13.24	13.34	13.28	13.37	13.54	13.43	13.35
13.52	13.23	13.48	13.48	13.54	13.41	13.51	13.44	13.36	13.36
13.53	13.44	13.69	13.66	13.32	13.26	13.51	13.38	13.46	13.34

Наблюдаем таблицу частот, в которой нам нужны столбцы **observed frequency** (наблюдаемые частоты) и **expected frequency** (ожидаемые частоты). Сравним графически наблюдаемые и ожидаемые частоты: выделим



соответствующие столбцы **Graphs** → **Custom Graphs** → **2D Graphs...** → **OK**. Наблюдаем некоторое различие.

В табл. 7.1 приведено значение статистики (7.8) **Chi-Square**: 12,00, количество степеней свободы **d.f.** = 3, которое получилось при объединении интервалов для выполнения условий (7.5):  $f = 6 - 1 - 2 = 3$ . Приведено значение вероятности

$$P\{\chi_3^2 \geq 12,00\} = p = 0,007.$$

Последнее означает, что если гипотеза верна, вероятность получить 12,00 или больше равна 0,007 – слишком мала, чтобы поверить в нормальность. Гипотезу о нормальности отклоняем.

Если посмотреть гистограмму наблюдений, видно, что в выборке имеется одно аномальное значение 14,56 (№ 188), которое могло появиться в результате какой-либо ошибки (при записи наблюдений, при перепечатке или попалась деталь с другого станка и т.д.). Удалим его и снова проверим гипотезу. Удаление одного наблюдения, если оно типично, не может изменить характеристики совокупности из 200 элементов; если же изменение происходит, следовательно, это наблюдение типичным не является и должно быть удалено.

Чтобы не портить исходные данные, продублируем их в новый столбец, например *dc*, и удалим аномальное наблюдение.

Повторим проверку гипотезы для «цензурированной» выборки и убедимся в том, что наблюдения не противоречат гипотезе о нормальности.

#### 4. Примеры проверки простой гипотезы о распределении

**Пример 7.2.** Проверим генератор случайных чисел. Сгенерируем выборку заданного объема с заданным в табл. 7.2 законом распределения, и по полученным результатам проверим гипотезу о согласии данных с этим распределением (файл с выборкой назовем, например, *Chisqr*). В табл. 7.2 приняты обозначения для распределений: *R* – равномерное, *N* – нормальное, *E* – показательное, *Bi* – биномиальное, *Po* – Пуассона.

#### Выполнение в пакете STATISTICA

Выполнение аналогично предыдущему.

Отличия от предыдущего: 1) в окне **Fitting Continuous Distribution** нужно ввести значения параметров распределения (вместо их оценок) и, возможно, поправить параметры группировки; 2) приводимый результат для уровня

значимости  $p$  не соответствует рассматриваемому случаю, так как число степеней свободы  $d.f.$  должно быть равным  $m - 1$ ; пакет же указывает с учетом числа оцениваемых параметров. Нужное значение для  $p$  получим в модуле **Basic Statistics and Tables** в *Probability calculator*.

**Таблица 7.2.** Исходные данные для примера 7.2

№ варианта	1	2	3	4	5	6
Распределение	$R[0, 5]$	$N(10, 2^2=4)$	$E(3)$	$Bi(10, 0.5)$	$Po(15)$	$beta(1, 1)$
Объем	130	140	140	160	130	140
№ варианта	7	8	9	10	11	12
Распределение	$R[0, 10]$	$N(15, 3^2=9)$	$E(5)$	$Bi(15, 0.3)$	$Po(20)$	$beta(2, 2)$
Объем	130	160	130	140	150	160
№ варианта	13	14	15			
Распределение	$R[-1, 1]$	$N(0, 1)$	$E(1)$			
Объем	130	140	150			

**Пример 7.3.** В опытах по генетике Мендель наблюдал частоты появления различных видов семян, получаемых при скрещивании гороха с круглыми желтыми и с морщинистыми зелеными семенами. Частоты приведены в табл. 7.3 вместе с теоретическими вероятностями.

**Таблица 7.3.** Частоты видов семян

Семена	Наблюдаемая частота, $v_i$	Теоретическая вероятность, $p_i$
Круглые и желтые	315	9/16
Морщинистые и желтые	101	3/16
Круглые и зеленые	108	3/16
Морщинистые и зеленые	32	1/16
Сумма	$n = 556$	

Формула (7.1) дает  $X^2 = 0.47$ . При числе степеней свободы  $m - 1 = 3$

$$P\{\chi_3^2 \geq 0.47\} = 0.92,$$

так что между наблюдениями и теорией имеется очень хорошее согласие: критерий с любым уровнем значимости  $\alpha \leq 0.92$  не отвергал бы эту гипотезу.

## Выполнение в пакете STATISTICA

Выполнить проверку гипотезы самостоятельно. Воспользоваться операциями со столбцами или процедурой *Observed versus expected* (наблюдаемые частоты против ожидаемых).

### 5. Проверка гипотезы о независимости признаков (таблица сопряженности признаков)

Предположим, имеется большая совокупность объектов, каждый из которых обладает двумя признаками  $A$  и  $B$ ; признак  $A$  имеет  $m$  уровней:  $A_1, \dots, A_m$ , а признак  $B$  –  $k$  уровней:  $B_1, \dots, B_k$ . Пусть уровень  $A_i$  встречается с вероятностью  $P(A_i)$ , а уровень  $B_j$  – с вероятностью  $P(B_j)$ . Признаки  $A$  и  $B$  независимы, если

$$P(A_i B_j) = P(A_i)P(B_j), \quad i = 1, \dots, m, \quad j = 1, \dots, k, \quad (7.10)$$

т.е. вероятность встретить комбинацию  $A_i B_j$  равна произведению вероятностей. Пусть признаки определены на  $n$  объектах, случайно извлеченных из совокупности;  $v_{ij}$  – число объектов, имеющих комбинацию  $A_i B_j$ ,  $\sum_{i=1}^m \sum_{j=1}^k v_{ij} = n$ . По

совокупности наблюдений  $\{v_{ij}\}$  (таблица  $m \times k$ ) требуется проверить гипотезу  $H$  о независимости признаков  $A$  и  $B$ . Задача сводится к случаю с неизвестными параметрами; ими являются вероятности

$$P(A_i), \quad i = 1, \dots, m; \quad P(B_j), \quad j = 1, \dots, k,$$

всего  $(m - 1) + (k - 1)$ ; их оценки:

$$\hat{P}(A_i) = \frac{\sum_{j=1}^k v_{ij}}{n} \equiv \frac{v_{i\cdot}}{n}, \quad \hat{P}(B_j) = \frac{\sum_{i=1}^m v_{ij}}{n} \equiv \frac{v_{\cdot j}}{n}$$

(в обозначениях точка означает суммирование по соответствующему индексу), и статистика (7.6) принимает вид

$$\tilde{X}^2 = \sum_{i=1}^m \sum_{j=1}^k \frac{v_{ij}^2}{n \hat{P}(A_i) \hat{P}(B_j)} - n = n \left( \sum_{i=1}^m \sum_{j=1}^k \frac{v_{ij}^2}{v_{i\cdot} v_{\cdot j}} - 1 \right). \quad (7.11)$$

Если гипотеза  $H$  верна, то по теореме Фишера  $\tilde{X}^2$  асимптотически распределена по закону хи-квадрат с числом степеней свободы

$$f = mk - 1 - (m - 1) - (k - 1) = (m - 1)(k - 1).$$

Поэтому, если

$$P\{\chi_f^2 \geq \tilde{X}^2\} \leq \alpha, \quad (7.12)$$

то гипотезу о независимости признаков следует отклонить.

Ясно, что по (7.11) – (7.12) можно проверять независимость двух случайных величин, разбив диапазоны их значений на  $m$  и  $k$  частей.

**Пример 7.4.** Данные, собранные по ряду школ, относительно физических недостатков школьников ( $P_1, P_2, P_3$  – признак  $A$ ) и дефектов речи ( $S_1, S_2, S_3$  – признак  $B$ ) приведены в табл. 7.4. В табл. 7.5 даны частоты.

**Таблица 7.4** Исходные данные

Дефекты речи (S) и физические недостатки (P) 217 школьников												
	P	S	P	S	P	S	P	S	P	S	P	S
1	P1	S1	P1	S1	P3	S2	P2	S2	P1	S3	P1	S1
2	P2	S3	P2	S2	P1	S3	P1	S1	P2	S2	P2	S1
3	P1	S1	P2	S3	P1	S2	P1	S1	P2	S2	P2	S2
4	P1	S2	P2	S3	P3	S1	P2	S1	P2	S2	P3	S3
5	P1	S1	P2	S1	P2	S1	P1	S1	P1	S1	P2	S1
6	P3	S1	P1	S2	P3	S3	P2	S2	P1	S3	P1	S1
7	P1	S1	P2	S3	P1	S2	P2	S2	P2	S1	P1	S2
8	P1	S2	P1	S1	P2	S3	P1	S2	P2	S2	P1	S3
9	P2	S2	P2	S1	P1	S2	P1	S1	P2	S2	P2	S3
10	P2	S2	P2	S1	P2	S2	P1	S3	P3	S3	P1	S1
11	P2	S2	P2	S1	P1	S2	P1	S2	P2	S1	P1	S1
12	P1	S2	P2	S2	P1	S2	P2	S2	P1	S1	P1	S1
13	P1	S1	P3	S3	P2	S2	P2	S2	P2	S1	P2	S3
14	P2	S3	P1	S1	P2	S3	P2	S1	P2	S1	P2	S1
15	P2	S1	P1	S1	P3	S2	P2	S2	P1	S1	P2	S2
16	P2	S1	P2	S1	P1	S2	P2	S1	P2	S2	P3	S3
17	P3	S2	P1	S1	P2	S2	P3	S3	P1	S1	P2	S1
18	P1	S1	P2	S2	P1	S1	P3	S2	P3	S3	P2	S2
19	P1	S2	P3	S3	P2	S1	P1	S1	P1	S1	P2	S2
20	P3	S3	P3	S3	P1	S1	P1	S1	P3	S2	P1	S1
21	P2	S2	P2	S1	P2	S3	P3	S2	P2	S2	P1	S2
22	P1	S3	P1	S1	P2	S2	P2	S2	P3	S1	P2	S2
23	P2	S3	P2	S2	P3	S3	P3	S3	P1	S1	P2	S1
24	P3	S2	P2	S2	P2	S3	P1	S3	P2	S2	P3	S2
25	P3	S1	P2	S3	P2	S1	P1	S2	P2	S2	P1	S2

Для проверки гипотезы о независимости этих двух признаков вычислим статистику (7.11):  $\tilde{X}^2 = 34,88$ ; число степеней свободы  $f = (3-1) \times (3-1) = 4$ ; минимальный уровень значимости

$$P\{\chi_4^2 \geq 34,88\} \leq 0,001.$$

Это значит, что при независимых признаках вероятность получить значение такое же, как в опыте или большее, меньше 0,001, и потому гипотезу о независимости следует отклонить.

### Выполнение в пакете STATISTICA

Образуем таблицу с двумя столбцами ( $P$  и  $S$ ) и 217 строками и назовем ее *Defects.sta* (это действие опускаем, если данные уже есть в компьютере). Работаем в модуле *Basic Statistics and Tables*:

*Analysis* → *Tables and banners* – в окне *Specify Table*, в поле *Analysis: Crosstabulation tables* кнопка *Specify Table* – отбираем признаки: *list 1: P, list 2: S* - **OK** - **OK** – в окне *Crosstabulation Tables Results* (результаты таблиц сопряженности) отмечаем (потребуем определить) *Expected frequencies* (ожидаемые или теоретические частоты) и *Pearson Chi-Square* → *Review Summary tables*.

**Таблица 7.5** Таблица частот

	$S_1$	$S_2$	$S_3$	Сумма
$P_1$	45	26	12	83
$P_2$	32	50	21	103
$P_3$	4	10	17	31
Сумма	81	86	50	217

Наблюдаем две таблицы: таблицу частот *Summary Frequency Table* и *Expected Frequencies*; в верхней части последней указано значение *Chi-square* статистики (7.11), число степеней свободы  $df$  и уровень значимости  $p$  (вероятность в (7.12)). Поскольку значение  $p$  мало, гипотеза о независимости речевых и физических дефектов отклоняется.

**Замечание 1.** Если бы исходные признаки  $X, Y, \dots$  были не символьными, а числовыми, нужно было бы сначала их классифицировать: разбить диапазон значений на части, и для каждой ввести свой символ (например,  $x_1, x_2, \dots, y_1, y_2, \dots$ ) введением дополнительных столбцов и использованием операции *Recode...* (кнопка *Vars* или *Edit* → *Variables*).

**Замечание 2.** Если бы исходными данными являлась таблица частот, то анализ можно было провести в модуле *Log - Linear Analysis* (как в п.6).

## 6. Проверка гипотезы об однородности выборок

Пусть имеется  $m$  выборок объемами  $n_1, \dots, n_m$ , извлеченных из различных совокупностей. Измеряемая величина в каждой из выборок может иметь  $k$  уровней  $B_1, \dots, B_k$ . Требуется проверить гипотезу о том, что исходные совокупности распределены одинаково. Обозначим  $v_{ij}$  – число наблюдений в  $i$ -й выборке, имеющих уровень  $B_j$ ,  $\sum_j v_{ij} \equiv v_{i\bullet} = n_i$ . Имеем таблицу  $m \times k$

наблюдений, аналогичную в предыдущем п. 5. Можно показать, что для проверки гипотезы справедлива процедура (7.11) – (7.12).

**Пример 7.5.** Имеются данные о наличии примесей серы в углеродистой стали, выплавляемой двумя заводами (табл. 7.6).

Проверим гипотезу о том, что распределения содержания серы (нежелательный фактор) одинаковы на этих заводах.

По (7.11) находим:  $\tilde{X}^2 = 3,39$ . Число степеней свободы  $f = (2-1) \times (4-1) = 3$ ; квантиль уровня 0,95

$$h = Q(0,95, 3) = 7,8.$$

Полученное нами из опыта значение 3,39 лежит в области допустимых значений, поэтому у нас нет оснований считать, что содержание серы в стали этих заводов имеют различные распределения.

**Таблица 7.6** Число плавов

	Содержание серы, $10^{-2}$ %				Сумма
	0÷2	2÷4	4÷6	6÷8	
Завод 1	82	535	1173	1714	3504
Завод 2	63	429	995	1307	2794
Сумма	145	964	2168	3021	

### Выполнение в пакете STATISTICA

Образуем таблицу  $2 \times 4$ , в которую занесем данные; столбцы назовем, например, S1 ÷ S4 (сера), а строки – Z1, Z2 (заводы). Работаем в модуле *Log - Linear Analysis*:

**Analysis** → **Startup Panel** – в поле **Input file: Frequencies w/out coding variables** (частоты без кодирующих переменных) – **Variables: Select All** → **OK** → **Specify table** (спецификация таблицы): **Factor Name: S, No. of levels** (число

уровней): 4, **Factor Name:** Z, **No. of levels:** 2 → OK – OK. В окне **Log - Linear Model Specification** выполним **Test all marginal**.

В таблице **Results of Fitting...** в последней строке столбца **Person Chi-Squ** получаем  $X^2 = 3,59$ , число степеней свободы **Degrs of Freedom f** = 3, и уровень значимости **Probab. p** = 0,31. Поскольку эта вероятность не мала (не является значимой), гипотезу об одинаковом распределении содержания серы в металле на двух заводах можно принять (вернее, наблюдения этому не противоречат).

### **Задание для самостоятельного изучения**

#### **Применение Т-критерия для сравнения средних в двух группах данных и F-теста для сравнения дисперсии**

На занятии 1 были изучены способы вычисления описательных статистик: выборочной средней, выборочной дисперсии, стандартного отклонения (выборочного СКО). Если имеются две группы данных, то естественно сравнить их путем сопоставления средних значений в этих группах.

#### **Цели задания**

**Знать** – *примеры ситуаций и задач, возникающих в ходе анализа данных, при решении которых требуется применение Т-критерия и F-критерия, ограничения и формальное определение указанных статистических критериев.*

**Уметь** – *осуществлять проверку гипотез сравнения средних и сравнения дисперсий двух выборок, делать выводы.*

**Владеть** – *навыками проверки статистических гипотез с использованием средств Excel и STATISTICA.*

#### **1. Т-критерий для сравнения средних в двух независимых выборках**

**Упражнение 1.** Используя учебники по теории вероятностей и математической статистике [4,11], изучите:

а) условия задач, при решении которых применяется Т-критерий сравнения средних для независимых выборок;

б) ограничения применения и формальное определение Т-критерия сравнения средних для независимых выборок.

**Упражнение 2.** Используя учебники по применению Excel для анализа данных [7,10], изучите:

а) возможности проверки гипотез о равенстве средних двух независимых выборок с помощью статистической функции

**TTEST(массив1; массив2; хвосты; 2)**

мастера функций  $f_x$  пакета Excel (для односторонней проверки гипотезы хвосты = 1, для двусторонней проверки хвосты = 2);

б) возможности проверки гипотез о равенстве средних двух независимых выборок с помощью надстройки Excel *Анализ данных*, выполняя работу следующим образом: *Данные* → *Анализ данных* → *Двухвыборочный t-тест с одинаковыми дисперсиями* → *ОК*.

Для испытания гипотез о равенстве средних в системе STATISTICA запустите модуль **Basic statistics/Tables (Основные статистики/таблицы)** из **Переключателя модулей**. В стартовой панели данного модуля выделите курсором строку **t-test for independent samples (t-тест для независимых выборок)** или **t-test for dependent samples (t-тест для зависимых выборок)** и нажмите кнопку **Ok** (рис. 7.1).

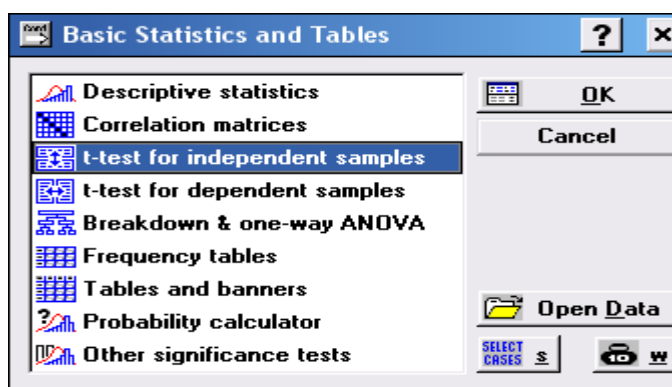


Рис. 7.1

**Упражнение 3.** Используя учебники по применению системы STATISTICA [3,14], изучите:

а) возможности проверки гипотез о равенстве средних двух независимых выборок по двум разным независимым выборкам (**t-test, independent, by variables**);

б) возможности проверки гипотез о равенстве средних двух независимых выборок для двух независимых групп, полученных из одной выборки при помощи группирующей переменной (**t-test, independent, by groups**).

## **2. Т-критерий для сравнения средних в двух зависимых выборках**



**Упражнение 4.** Используя учебники по теории вероятностей и математической статистике [4,11], изучите условия задач, при решении которых применяется Т-критерий сравнения средних для двух зависимых выборок; ограничения применения и формальное определение Т-критерия сравнения средних для двух зависимых выборок.

**Упражнение 5.** Используя учебники по применению Excel для анализа данных [7,10], изучите возможности проверки гипотез о равенстве средних двух зависимых выборок с помощью статистической функции

**TTEST(массив1; массив2; хвосты; 1)**

мастера функций  $f_x$  пакета Excel (для односторонней проверки гипотезы хвосты = 1, для двусторонней проверки хвосты = 2) и с помощью надстройки Excel *Анализ данных*, выполняя работу следующим образом: *Данные* → *Анализ данных* → *Парный двухвыборочный t-тест для средних* → *ОК*.

**Упражнение 6.** Используя учебники по применению системы STATISTICA [3,14], изучите возможности проверки гипотез о равенстве средних двух зависимых выборок (**t-test, dependent samples**).

### **3. Двухвыборочный F-тест для дисперсии**

В ряде задач для сравнения двух независимых выборок необходимо проверить гипотезу о равенстве дисперсий (в частности, равенство дисперсий является одним из условий для применения t-критерия о равенстве средних).

**Упражнение 7.** Используя учебники по теории вероятностей и математической статистике [4,11], изучите условия задач, при решении которых применяется F-тест для проверки гипотез о равенстве дисперсий двух генеральных совокупностей; ограничения применения и формальное определение F-критерия.

**Упражнение 8.** Используя учебники по применению Excel для анализа данных [7,10], изучите возможности проверки гипотез о равенстве дисперсий двух зависимых выборок с помощью статистической функции

**FTEST(массив1; массив2)**

мастера функций  $f_x$  пакета Excel и с помощью надстройки Excel *Анализ данных*, выполняя работу следующим образом: *Данные* → *Анализ данных* → *Двухвыборочный f-тест для дисперсии* → *ОК*.

**Замечание.** В системе STATISTICA гипотеза о равенстве дисперсий двух выборок проверяется одновременно с проверкой гипотез о равенстве средних. Результаты расчета соответствующих значений Т- и F-статистик приводятся в одной итоговой таблице.

## **Занятие 8**

### **Проверка гипотезы о виде распределения и однородности**

#### **выборок с помощью критерия Пирсона. Лабораторная работа № 4**

Гипотеза о виде распределения обычно возникает на основании результатов разведочного анализа: изучения гистограммы, асимметрии, эксцесса. При этом параметры распределений заменяются их выборочными оценками.

Проверка гипотезы об однородности нужна потому, что однородность является важнейшим качеством выборки: это свидетельство тому, что измерения проводились в стабильных условиях или что выборка извлечена из одной генеральной совокупности.

### **Цель занятия**

*Формирование навыков анализа данных с использованием возможностей Excel и STATISTICA.*

#### **Лабораторная работа № 4 Проверка статистических гипотез.**

Выполнение лабораторной работы разобьем на три этапа.

#### ***Порядок выполнения работы на первом этапе***

**Цель работы:** проверить гипотезу о виде распределения величины  $X$ , используя критерий Пирсона при уровне значимости 0,05 (либо 0,01 при неудаче)

1. Сгенерировать 5 выборок.
2. Вычислить числовые характеристики полученных выборок: выборочное среднее, выборочную дисперсию, моду, медиану. Для расчета использовать Excel, Statistica.
3. Сравнить полученные результаты с теоретическими значениями числовых характеристик для соответствующих случайных величин.
4. Построить гистограммы относительных частот и графики эмпирических функций распределения.

5. Проверить гипотезу о виде распределения.

**Таблица 8.1** Варианты заданий

№ варианта	1	2	3	4	5
Распределение	$N(0, 1)$	$R[-1, 1]$	$E(3)$	$N(10, 4)$	$R[0, 10]$
Объем	80	100	150	80	100
№ варианта	6	7	8	9	10
Распределение	$E(5)$	$N(15, 4)$	$R[5, 15]$	$E(10)$	$R[20, 25]$
Объем	150	80	100	150	80
№ варианта	11	12	13	14	15
Распределение	$N(1, 1)$	$R[-2, 2]$	$E(0,5)$	$N(-1, 1)$	$R[0, 5]$
Объем	80	100	150	80	100
№ варианта	16	17	18	19	20
Распределение	$E(3)$	$N(10, 4)$	$R[5, 10]$	$E(6)$	$R[15, 25]$
Объем	150	80	100	150	100

***Порядок выполнения работы на втором этапе***

Цель работы: проверить гипотезу об однородности трех выборок.

Сгенерировать три выборки объемами  $n_1 = 180$ ,  $n_2 = 100$ ,  $n_3 = 120$  для заданного в табл. 8.2 распределения.

Провести их группирование на  $8 \div 10$  интервалах.

Проверку гипотез об однородности выборок выполнить для 2-х вариантов:

а) параметры одинаковы; б) параметры различны.

**Таблица 8.2** Исходные данные.

N	Тип	Вариант1	Вариант 2		
		$a_1 = a_2 = a_3$	$a_1$	$a_2$	$a_3$
1	$N(a, 1)$	10	9.8	10	11.2
2	$E(a)$	10	8.0	10	12.0
3	$Po(a)$	10	9.5	10	11.5
4	$N(a, 2)$	20	19.5	20	21.5
5	$E(a)$	20	16.0	20	24.0
6	$Po(a)$	20	19.0	20	21.0
7	$N(a, 3)$	30	29.4	30	30.6
8	$E(a)$	30	24.0	30	36.0
9	$Po(a)$	30	28.0	30	32.0
10	$N(a, 4)$	40	39.0	40	41.0

**Выполнение в пакете STATISTICA**

Группирование провести процедурой *Frequency tables*, и из трех таблиц сформировать одну. Гипотезу об однородности проверить аналогично п.6.

**Отчет по работе** должен содержать:

- 1) краткое описание критерия  $\chi^2$ -квадрат;
- 2) постановки конкретных задач;
- 3) несколько значений анализируемых выборок;
- 4) сгруппированные данные;
- 5) результаты основных вычислений и статистические выводы.

### ***Порядок выполнения работы на третьем этапе***

Цель работы: проверить гипотезу о виде распределения величины  $X$ , используя критерий Пирсона при уровне значимости 0,05.

1. Найти реальные данные, включающие не менее 100 значений некоторой случайной величины. Обязательно привести доказательства их реальности (например, Интернет-ссылка, копия журнала, подпись ответственного лица и т. п.).

2. По данным построить гистограмму и выдвинуть гипотезу о виде распределения<sup>1</sup> (нормальное, экспоненциальное, равномерное или др.).

3. Провести необходимые вычисления для проверки гипотезы.

4. Сделать выводы по работе.

*При сдаче лабораторной работы демонстрируется:*

- 1) источник, откуда взяты данные;
- 2) гистограмма и все вычисления в электронном, либо бумажном виде;
- 3) качественный отчёт.

### **Задание для самостоятельного изучения**

#### **Корреляционный анализ**

На практике часто встречаются случайные величины, возможные значения которых определяются несколькими числами. Двумерной называют случайную величину, значения которой есть пары чисел  $(x, y)$ . Геометрически ее можно истолковать как случайную точку  $M(X, Y)$  на плоскости  $xOy$  либо как случайный вектор  $OM$ .

Согласно [3], цель всякого исследования или научного анализа состоит в нахождении связей (зависимостей) между переменными и не существует иного

---

<sup>1</sup> Будет считаться ошибкой выдвижение гипотезы, например, о нормальном распределении при виде гистограммы явно подходящим под другое распределение.

способа представления знания, кроме как в терминах зависимостей между количествами и качествами.

### Цели задания

**Знать** – определения и формулы вычисления ковариации и коэффициента корреляции, а также формулы вычисления их статистических оценок.

**Уметь** – вычислять значения выборочной ковариации и выборочного коэффициента корреляции, проверять статистическую гипотезу о значимости коэффициента корреляции.

**Владеть** – методами корреляционного анализа с использованием средств Excel и STATISTICA.

### Числовые характеристики двумерных случайных величин

**1. Корреляционным моментом (ковариацией)** случайных величин  $X$  и  $Y$  называют математическое ожидание произведения отклонений этих величин от своих математических ожиданий:  $\mu_{xy} = M[(X - M(X))(Y - M(Y))]$ .

Для вычисления корреляционного момента дискретных случайных величин используют формулу  $\mu_{xy} = \sum_{i=1}^n \sum_{j=1}^m (x_i - M(X))(y_j - M(Y))p(x_i | y_j)$ .

Для непрерывных случайных величин:

$$\mu_{xy} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - M(X))(y - M(Y))f(x, y)dx dy.$$

Корреляционный момент служит для характеристики связи между величинами  $X$  и  $Y$ . Если  $X$  и  $Y$  независимы, корреляционный момент равен 0, иначе корреляционный момент отличен от нуля.

**2. Коэффициентом корреляции** ( $\rho_{xy}$ ) случайных величин  $X$  и  $Y$  называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:  $\rho_{xy} = \frac{\mu_{xy}}{\sigma_X \sigma_Y}$ .

Напомним, что коэффициент корреляции – безразмерная величина, причем  $|\rho_{xy}| \leq 1$ . Он служит для оценки тесноты линейной связи между  $X$  и  $Y$ : чем ближе абсолютная величина коэффициента корреляции к единице, тем связь сильнее; чем ближе абсолютная величина коэффициента корреляции к нулю, тем связь слабее. **Коррелированными** называют две случайные величины, если их корреляционный момент отличен от нуля. **Некоррелированными** называют две случайные величины, если их корреляционный момент равен нулю. Из

независимости двух случайных величин следует их некоррелированность, обратное, вообще говоря, неверно.

### Точечные оценки ковариации и коэффициента корреляции

Пусть имеется выборка реализаций двумерной случайной величины  $(X, Y)$ , т.е.  $n$  пар чисел  $(X_i, Y_i)$ ,  $i = 1, \dots, n$ .

Выборочную характеристику  $K_{xy} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (x_i - \bar{x})(y_j - \bar{y})$  называют

**корреляционным моментом**, или **выборочной ковариацией**. Величина  $K_{xy}$  является точечной оценкой ковариации случайных величин  $X$  и  $Y$ .

Выборочную характеристику  $r_{xy} = \frac{K_{xy}}{\hat{\sigma}_x \hat{\sigma}_y}$  называют **выборочным коэффициентом корреляции**. Величина  $r_{xy}$  является точечной оценкой  $\rho_{xy}$  – коэффициента корреляции случайных величин  $X$  и  $Y$ .

**Замечание.** Точечные оценки математического ожидания  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  и  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  (выборочные средние) и точечные оценки дисперсии  $\hat{\sigma}_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$  и  $\hat{\sigma}_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$  (выборочные дисперсии) рассматривались на занятии 1.

### Вычисление точечных оценок ковариации и коэффициента корреляции в Excel

**А) По формулам.** Рассмотрим две выборки размера  $n_1 = n_2 = 10$ .  $X$  – общие коэффициенты рождаемости и  $Y$  – среднедушевые денежные доходы населения в регионах Центральной России (данные взяты с сайта федеральной службы государственной статистики [www.gks.ru](http://www.gks.ru)). Покажем, как с помощью указанных выше формул вычислить в Excel ковариацию и коэффициент корреляции.

Введём исходные данные в электронную таблицу Excel (столбцы А, В и С на рис. 8.1). Вычислим в столбцах D и E выборочные средние  $\bar{x}$  и  $\bar{y}$  для  $X$  и  $Y$  соответственно: =СУММ(B2:B11)/10 и =СУММ(C2:C11)/10. Далее в столбцах F и G находим отклонение вариантов  $x_i$  и  $y_i$  от выборочных средних  $\bar{x}$  и  $\bar{y}$ , а в

столбце Н – их произведение. Тогда выборочную ковариацию  $K_{xy}$  в столбце I можно вычислить как: =СУММ(Н2:Н11)/10. Для определения выборочного коэффициента корреляции  $r_{xy}$  вычислим сначала в столбцах J и K средние квадратические отклонения величин **X** и **Y**, воспользовавшись функцией **СТАНДОТКЛОНП**, то есть: =СТАНДОТКЛОНП(В2:В11) для нахождения  $\hat{\sigma}_x$  и =СТАНДОТКЛОНП(С2:С11) для нахождения  $\hat{\sigma}_y$ . И тогда  $r_{xy}$  вычисляем в столбце L как: =I2/(J2\*K2).

	A	B	C	D	E	F	G	H	I	J	K	L
1		Общ. коэфф. рожд. $x_i$	Среднедуш. доход $y_i$	$\bar{x}$	$\bar{y}$	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$\mu_{xy}$	$\sigma_x$	$\sigma_y$	$r_{xy}$
2	Белгородская обл.	11,0	14117	10,45	12351,7	0,55	1765,3	970,915	58,55	0,57	1558,06	0,065
3	Брянская обл.	11,1	11404			0,65	-947,7	-616,005				
4	Воронежская обл.	10,4	11728			-0,05	-623,7	31,185				
5	Ивановская обл.	10,6	9343			0,15	-3008,7	-451,305				
6	Калужская обл.	10,5	13509			0,05	1157,3	57,865				
7	Курская обл.	10,8	12634			0,35	282,3	98,805				
8	Липецкая обл.	10,9	14686			0,45	2334,3	1050,435				
9	Орловская обл.	10,4	10660			-0,05	-1691,7	84,585				
10	Тамбовская обл.	9,3	12077			-1,15	-274,7	315,905				
11	Тульская обл.	9,5	13359			-0,95	1007,3	-956,935				

Рис. 8.1

Найденное значение выборочного коэффициента  $r_{xy} = 0,065$  близко к нулю, таким образом, между общим коэффициентом рождаемости (выборка **X**) и среднедушевым денежным доходом (выборка **Y**) в регионах Центральной России существует лишь слабая корреляция, что может показаться странным на первый взгляд.

**Б) С помощью статистических функций.** Как было сказано в занятии 1, в Excel представлен широкий набор встроенных статистических функций. Так, для вычисления точечных оценок ковариации  $\mu_{xy}$  и коэффициента корреляции  $\rho_{xy}$  выборок **X** и **Y** можно воспользоваться функциями **КОВАР** и **КОРРЕЛ** соответственно. В качестве аргументов в обоих случаях следует задавать массивы из выборок, например, для предыдущего примера (рис. 8.1) это реализуется так: =КОВАР(В2:В11;С2:С11) и =КОРРЕЛ(В2:В11;С2:С11).

Заметим, что если в качестве аргументов функции **КОВАР** задать один и тот же массив данных, то мы получим значение дисперсии этой случайной величины, обычно вычисляемой в Excel с помощью функции **ДИСПР**.

**В) С помощью пакета Анализ данных.** Использовать надстройку «Пакет анализа» при проведении ковариационного или корреляционного анализа имеет смысл, если рассматривается более двух случайных величин. Дополним наши

данные выборкой **Z** – долей сельского населения в регионах (столбец D на рис. 8.3).

Приступим к ковариационному анализу: *Данные* → *Анализ данных* → *Ковариация* → *ОК* (рис. 8.2).

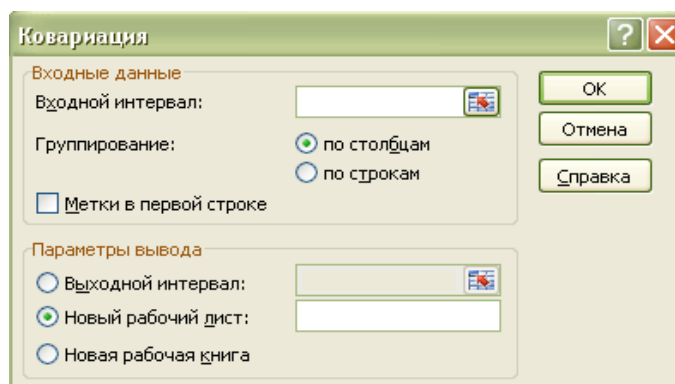


Рис. 8.2

В появившемся диалоговом окне в качестве входного интервала выбираем диапазон смежных столбцов, содержащих анализируемые данные, а в качестве выходного интервала вводим ссылку на левую верхнюю ячейку ковариационной таблицы (или можно выбрать новый рабочий лист) и нажимаем **ОК**. Результат ковариационного анализа представлен на рис. 8.3.

	A	B	C	D	E	F	G	H
1		Общ. коэфф. рожд. $x_i$	Среднедуш. доход $y_i$	Доля сельск. населения $z_i$		$X$	$Y$	$Z$
2	Белгородская обл.	11,0	14117	33,5	$X$	0,33		
3	Брянская обл.	11,1	11404	31,4	$Y$	58,55	2427565,21	
4	Воронежская обл.	10,4	11728	36,5	$Z$	0,03	1759,02	52,83
5	Ивановская обл.	10,6	9343	19,3				
6	Калужская обл.	10,5	13509	23,7				
7	Курская обл.	10,8	12634	35,3				
8	Липецкая обл.	10,9	14686	35,7				
9	Орловская обл.	10,4	10660	35,7				
10	Тамбовская обл.	9,3	12077	42,0				
11	Тульская обл.	9,5	13359	20,1				

Рис. 8.3

Таким образом, ковариационный анализ в пакете *Анализ данных* вычисляет значение функции ковариации для всех возможных пар случайных величин, при этом элементы, находящиеся на диагонали ковариационной таблицы, не что иное, как дисперсии величин **X**, **Y** и **Z**.

Аналогичным образом проводится и корреляционный анализ: *Данные* → *Анализ данных* → *Корреляция* → *ОК*. Результаты представлены на рис. 8.4.



	A	B	C	D	E	F	G	H
1		Общ. коэфф. рожд. $x_i$	Среднедуш. доход $y_i$	Доля сельск. населения $z_i$		$X$	$Y$	$Z$
2	Белгородская обл.	11,0	14117	33,5	$X$	1,000	0,065	0,008
3	Брянская обл.	11,1	11404	31,4	$Y$	0,065	1,000	0,155
4	Воронежская обл.	10,4	11728	36,5	$Z$	0,008	0,155	1,000
5	Ивановская обл.	10,6	9343	19,3				
6	Калужская обл.	10,5	13509	23,7				
7	Курская обл.	10,8	12634	35,3				
8	Липецкая обл.	10,9	14686	35,7				
9	Орловская обл.	10,4	10660	35,7				
10	Тамбовская обл.	9,3	12077	42,0				
11	Тульская обл.	9,5	13359	20,1				

Рис. 8.4

Попробуйте самостоятельно прокомментировать результаты корреляционного анализа.

### Построение корреляционных матриц в системе STATISTICA

Для построения корреляционной матрицы в системе STATISTICA необходимо осуществить следующие переходы: **Статистика** → **Основная статистика/Таблицы** → **Correlation Matrices** → **ОК**. Откроется рабочее окно модуля (рис. 8.5).

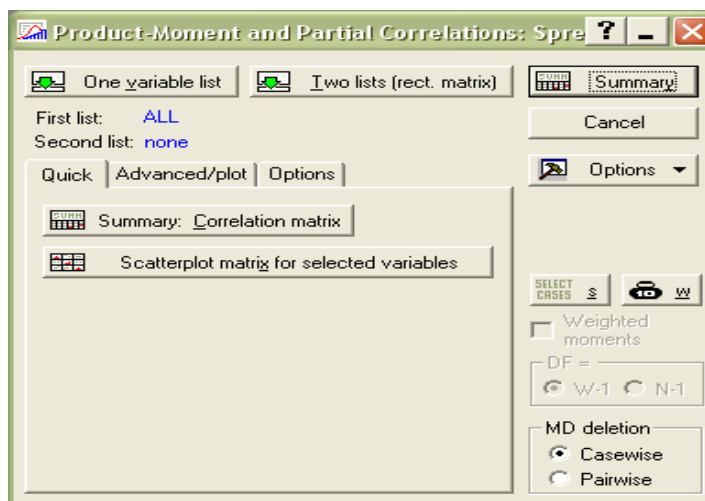


Рис. 8.5

Имена переменных нужно задать одним списком (кнопка **One variables list**). В этом случае после нажатия кнопки **Summary. Correlation Matrix** будет построена квадратная корреляционная матрица, строки и столбцы которой представлены списком переменных (рис. 8.6). Элементы матрицы —

коэффициенты корреляции между переменными, расположенными на пересечении строки и столбца.

Кнопка *Scatterplot matrix for selected variables* позволяет построить графики функции рассеяния и гистограммы выбранных переменных.

Вкладка *Advanced/plot* предоставляет расширенные услуги графической иллюстрации статистического анализа выделенных переменных.

Variable	X	Y	Z
X	1,000	0,065	0,008
Y	0,065	1,000	0,155
Z	0,008	0,155	1,000

Рис. 8.6

На вкладке *Options* можно изменить параметры корреляционного анализа. Если установить флажок на *Display detailed table of results*, то наряду с коэффициентами корреляции будут даны результаты статистического анализа переменных: средние; стандартные отклонения; значения t-критерия сравнения средних и др.

### Проверка гипотезы о значимости коэффициента корреляции

Пусть двумерная величина  $(X,Y)$  распределена нормально. Для выборки объема  $n$  найдем выборочный коэффициент корреляции  $r_{xy}$ . Требуется проверить гипотезу о равенстве нулю коэффициента корреляции генеральной совокупности, т.е. нулевую гипотезу  $H_0: \rho_{xy}=0$ . Если нулевая гипотеза принимается, то это означает, что  $X$  и  $Y$  – некоррелированы; в противном случае – коррелированы.

Для того чтобы при уровне значимости  $\alpha$  проверить нулевую гипотезу о равенстве нулю коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе  $H_1: \rho_{xy} \neq 0$ , надо вычислить наблюдаемое

значение критерия  $T_{набл} = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$  и по таблице критических точек распределения

Стьюдента по заданному уровню значимости  $\alpha$  и числу степеней свободы  $k = n - 2$  найти критическую точку  $t_{кр}(\alpha, k)$  двусторонней критической области.

Если  $|T_{набл}| > t_{кр}$ , то нулевую гипотезу отвергают. Если  $|T_{набл}| < t_{кр}$  – нет оснований отвергать нулевую гипотезу.

**Замечание.** Найти критическое значение распределения Стьюдента можно с помощью таблиц в учебниках по теории вероятностей и математической

статистике [4,13], в Excel – с помощью статистической функции **СТЮДЕНТОБР**, в STATISTICA – с помощью средства **Вероятностный калькулятор**.

**Замечание.** При вычислении корреляционных матриц в системе STATISTICA значимые коэффициенты корреляции выделяются красным цветом.

**Пример.** Выше при рассмотрении двух выборок размера  $n_1 = n_2 = 10$  **X** (общие коэффициенты рождаемости) и **Y** (среднедушевые денежные доходы населения в регионах Центральной России) было получено выборочное значение коэффициента корреляции  $r_{xy} = 0,065$ . Так как полученное значение не велико, естественно выдвинуть гипотезу об отсутствии линейной связи между данными показателями, т.е.  $H_0: \rho_{xy} = 0$  при альтернативной гипотезе  $H_1: \rho_{xy} \neq 0$ . Уровень значимости равен 0,05.

Так как объем выборки  $n=10$ , наблюдаемое значение критерия  $T_{набл} = \frac{0,065\sqrt{10-2}}{\sqrt{1-0,065^2}} = 0,184$ . Количество степеней свободы  $k = 8$ . Критическое значение  $t_{кр} = t_{кр}(0,05; 8)$  найдем в Excel с помощью статистической функции **=СТЮДРАСПОБР(0,025; 8)**, где первый параметр равен 0,05/2 (для двусторонней критической области), второй параметр равен количеству степеней свободы. В результате получено значение  $t_{кр} = 2,75$ .

Так как  $|T_{набл}| < t_{кр}$ , нет оснований отвергать нулевую гипотезу, т.е. между показателями **X** и **Y** статистически значимая линейная связь отсутствует.

### **Задание для самостоятельной работы**

Найти данные (в Интернете, журналах, статистических сборниках, справочниках). Используя средства Excel и STATISTICA, вычислить выборочные коэффициенты корреляции и сделать вывод о тесноте и направлении связей между выбранными показателями.

## ПРИЛОЖЕНИЯ

**Таблица П1** Индивидуальные задания по вариантам для выполнения лабораторной работы № 1

№	Опыт	Событие	Количество серий	Значение инкремента
1	Бросание двух костей	Набор в сумме 8 очков	1800	30
2	Вытягивание одного шара из 4 белых и 5 чёрных шаров	Вытягивание чёрного шара	1200	20
3	Вытягивание 2 карт из 36-картовой колоды	Карты одной масти	1500	50
4	Вытягивание карты из 36-картовой колоды	Вытягивание карты червовой масти	1300	25
5	Бросание кости	Выпадение «6»	1400	25
6	Вытягивание 3 карт из 36-картовой колоды	Карты разных мастей	1900	38
7	Вытягивание 2 карт из 36-картовой колоды	Обе карты старше валета	2000	50
8	Вытягивание 2 карточек с цифрами (10 карточек)	Если цифры сложить – получится чётное число	1500	30
9	Вытягивание одного шара из 11 пронумерованных шаров	Вытягивание шара № 5	1400	20
10	Бросание кости	Выпадение «3» или «4»	1200	25
11	Вытягивание 2 карточек с цифрами (10 карточек)	Если цифры сложить – получится нечётное число	1600	40
12	Вытягивание карты из 52-картовой колоды	Вытягивание любой дамы	1800	50
13	Генерация ЭВМ трёхзначного числа (от 100 до 999)	Число содержит хотя бы одну 3	2000	40
14	Генерация ЭВМ трёхзначного числа (от 100 до 999)	Число кратно 7	2000	50
15	Бросание двух костей	Набор в сумме 10 очков	1000	20

**Окончание табл. П1**

№	Опыт	Событие	Количество серий	Значение инкремента
16	Вытягивание карты из 36-картовой колоды	Вытягивание червовой масти	1500	15
17	Бросание кости	Выпадение «1» или «2»	1000	20
18	Игра в лото (90 бочонков)	Вытягивание бочонка с простым числом	1200	40

19	Вытягивание карты из 36-картовой колоды	Вытягивание туза	2000	25
20	Бросание двух костей	Набор в сумме 9 очков	1400	20
21	Игра в домино (28 костей)	Вытягивание кости с дублем	1000	20
22	Игра в лото (90 бочонков)	Вытягивание бочонка с составным числом	900	30
23	Вытягивание карты из 36-картовой колоды	Вытягивание червовой дамы	2500	50
24	Вытягивание одного шара из 3 белых и 5 чёрных шаров	Вытягивание белого шара	1200	25
25	Вытягивание одного шара из 8 пронумерованных шаров	Вытягивание шара № 3	700	20
26	Игра в домино (28 костей)	Вытягивание костяшки без дубля	1300	13
27	Вытягивание карты из 52-картовой колоды	Вытягивание любого туза	1700	34
28	Вытягивание одного шара из 5 белых и 7 чёрных шаров	Выпадение белого шара	800	20
29	Вытягивание карточки с буквами (33 буквы)	Выпадение гласной буквы	1100	22
30	Вытягивание 2 карт из 36-картовой колоды	Вторая карта старше первой	1500	30

**Таблица П2** Индивидуальные задания по вариантам для выполнения лабораторной работы № 2

№	Обязательные условия для выбираемых функций
1	В выбранной функции должны присутствовать $\sin$ и $\ln$
2	В выбранной функции должны присутствовать $sh$ и $cos$
3	В выбранной функции должны присутствовать $sec$ и $a^x$ , где $a > 1$ любое
4	В выбранной функции должны присутствовать $\arcsin$ и $x^a$ , где $a > 1$ любое
5	В выбранной функции должны присутствовать $th$ и $x^a$ , где $a > 1$ любое
6	В выбранной функции должны присутствовать $tg$ и $a^x$ , где $a > 1$ любое
7	В выбранной функции должны присутствовать $ch$ и $arctg$
8	В выбранной функции должны присутствовать $ctg$ и $a^x$ , где $a > 1$ любое
9	В выбранной функции должны присутствовать $cosec$ и $\ln$
10	В выбранной функции должны присутствовать $\sin$ и $x^x$
11	В выбранной функции должны присутствовать $\arccos$ и $x^a$ , где $a > 1$ любое
12	В выбранной функции должны присутствовать $cth$ и $\log_a$ , где $a > 1$ любое
13	В выбранной функции должны присутствовать $arcsh$

14	В выбранной функции должны присутствовать $\ln$ и $x^a$ , где $a > 1$ любое
15	В выбранной функции должны присутствовать $\cos$ и $x^a$ , где $a > 1$ любое

### **Библиографический список**

1. Белова, И.М. Компьютерное моделирование / И.М. Белова. – М.: МГИУ, 2008. – 81 с.
2. Боровиков, В.П. Программа STATISTICA для студентов и инженеров / В.П. Боровиков. – М.: КомпьютерПресс, 2001. – 301 с.
3. Боровиков, В.П. STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В.П. Боровиков. – СПб.: Питер, 2001. – 656 с.
4. Гмурман, В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. – М.: Высшее образование, 2006. – 479 с.
5. Ермаков, С.М. Статистическое моделирование / С.М. Ермаков, Г.А. Михайлов. – М.: Наука, 1982. – 296 с.
6. Ермаков, С.М. Метод Монте-Карло в вычислительной математике: ввод. курс / С.М. Ермаков. - М.; СПб.: БЛЗ: Нев. диалект, 2009. – 192 с.
7. Макарова, Н.В. Статистика в Excel / Н.В. Макарова, В.Я. Трофимец. – М.: Финансы и статистика, 2006. – 368 с.
8. Математическая статистика: учеб. для вузов / под ред. В.С. Зарубина, А.В. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2002. – 424 с.
9. Михайлов, Г.А. Численное статистическое моделирование: методы Монте-Карло / Г.А. Михайлов, А.В. Войтишек. – М.: Академия, 2006. – 366 с.
10. Просветов, Г.И. Анализ данных с помощью Excel: Задачи и решения / Г.И. Просветов. – М.: Альфа-Пресс, 2009. – 255 с.
11. Просветов, Г.И. Теория вероятностей и математическая статистика: Задачи и решения / Г.И. Просветов. – М.: Альфа-Пресс, 2009. – 272 с.
12. Соболев, И.М. Метод Монте-Карло / И.М. Соболев. – М.: Наука, 1978. – 64 с.
13. Теория вероятностей: учеб. для вузов / под ред. В.С. Зарубина, А.В. Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2004. – 456 с.
14. Халафян, А. А. STATISTICA 6. Статистический анализ данных / А.А. Халафян. – М.: Бином, 2007. – 512 с.