

МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ МЕДИЦИНСКИЙ УНИВЕРСИТЕТ
КАФЕДРА МЕДИЦИНСКОЙ И БИОЛОГИЧЕСКОЙ ФИЗИКИ

М. А. ШЕЛАМОВА, Н. И. ИНСАРОВА, В. Г. ЛЕЩЕНКО

**ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА
МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ
С ИСПОЛЬЗОВАНИЕМ ПРОГРАММЫ EXCEL**

Учебно-методическое пособие



Минск БГМУ 2017

УДК 577.3:004(075.8)

ББК 32.97я73

Ш42

Рекомендовано Научно-методическим советом университета в качестве учебно-методического пособия 19.04.2017 г., протокол № 8

Рецензенты: канд. мед. наук, доц. О. Н. Ринейская; канд. хим. наук, доц. С. В. Ткачев

Шеламова, М. А.

Ш42 Основы статистического анализа медико-биологических данных с использованием программы Excel : учебно-методическое пособие / М. А. Шеламова, Н. И. Инсарова, В. Г. Лещенко. – Минск : БГМУ, 2017. – 92 с.

ISBN 978-985-567-841-1.

Рассмотрены основные понятия теории вероятностей и математической статистики, необходимые для проведения статистического анализа результатов медико-биологических исследований. Многочисленные примеры демонстрируют порядок проведения соответствующих расчетов на компьютере с использованием табличного процессора Excel.

Предназначено для студентов всех факультетов, аспирантов и магистрантов.

УДК 577.3:004(075.8)

ББК 32.97я73

Учебное издание

Шеламова Марина Алексеевна
Инсарова Наталия Васильевна
Лещенко Вячеслав Григорьевич

**ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА
МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ
ПРОГРАММЫ EXCEL**

Учебно-методическое пособие

Ответственный за выпуск М. В. Гольцев
Компьютерная верстка Н. М. Федорцовой

Подписано в печать 11.10.17. Формат 60×84/16. Бумага писчая «Снегурочка».

Ризография. Гарнитура «Times».

Усл. печ. л. 5,35. Уч.-изд. л. 4,5. Тираж 225 экз. Заказ 698.

Издатель и полиграфическое исполнение: учреждение образования
«Белорусский государственный медицинский университет».

Свидетельство о государственной регистрации издателя, изготовителя,
распространителя печатных изданий № 1/187 от 18.02.2014.

Ул. Ленинградская, 6, 220006, Минск.

ISBN 978-985-567-841-1

© Шеламова М. А., Инсарова Н. И., Лещенко В. Г., 2017

© УО «Белорусский государственный медицинский университет», 2017

ВВЕДЕНИЕ

Математико-статистическое описание данных медицинских исследований и оценка значимости различия величин, характеризующих эффективность проводимых профилактических, диагностических и лечебных мероприятий, являются основополагающими для доказательной медицины.

Не отягощая читателя излишней математикой, авторы учебно-методического пособия пытались решить две задачи:

1. Определить базовые понятия теории вероятностей и статистики и объяснить их смысл. Без этого невозможно осмысленно применять методы статистического анализа данных (гл. 1 и 2).

2. Рассмотреть этапы проведения анализа на конкретных примерах, используя табличный процессор Excel таким образом, чтобы любой студент-медик, аспирант или даже врач, взяв это учебно-методическое пособие и определив цель исследования, всегда мог самостоятельно получить нужный результат (гл. 3).

В издании рассмотрены далеко не все используемые сегодня на практике статистические методы. Мы ограничиваемся следующими: «Описательная статистика», «Элементы корреляционного анализа», «Оценка значимости различия признаков (статистические гипотезы и критерии проверки гипотез)». Именно они, как показывает знакомство с медицинской литературой, прежде всего необходимы при анализе полученных результатов. Ограничение так же связано с малым временем, отведенным программой на эту работу.

Глава 1

ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ЗАКОНЫ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

1.1. ЗАКОНОМЕРНОСТЬ И СЛУЧАЙНОСТЬ, СЛУЧАЙНАЯ ИЗМЕНЧИВОСТЬ В ТОЧНЫХ НАУКАХ, БИОЛОГИИ И МЕДИЦИНЕ

Теория вероятностей — область математики, которая изучает закономерности в *случайных явлениях*. Случайное явление — это явление, которое при неоднократном воспроизведении одного и того же опыта может протекать каждый раз несколько по-иному.

Очевидно, что в природе нет ни одного явления, в котором не присутствовали бы в той или иной мере элементы случайности, но в различных ситуациях мы учитываем их по-разному. Так, в ряде практических задач ими можно пренебречь и рассматривать вместо реального явления его упрощенную схему — «модель», предполагая, что в данных условиях опыта оно протекает вполне определенным образом. При этом выделяются самые главные, решающие факторы, характеризующие явление. Именно такая схема изучения явлений чаще всего применяется в физике и технике.

Однако при решении многих задач многочисленные, тесно переплетающиеся между собой случайные факторы часто играют определяющую роль. Здесь на первый план выступает *случайная природа* явления, которой уже нельзя пренебречь. Это явление необходимо изучать именно с точки зрения закономерностей, присущих ему как случайному явлению.

Предмет изучения биологов и медиков — живой организм, зарождение, развитие и существование которого определяется очень многими и разнообразными, часто случайными внешними и внутренними условиями. Именно поэтому явления и события живого мира во многом тоже *случайны* по своей природе.

Элементы неопределенности, сложности, многопричинности, присущие случайным явлениям, обуславливают необходимость создания специальных математических методов для их изучения.

Разработка таких методов, установление специфических закономерностей, свойственных случайным явлениям — главные задачи теории вероятностей. Характерно, что эти закономерности выполняются лишь при массовости случайных явлений. Причем индивидуальные особенности отдельных случаев как бы взаимно погашаются, а усредненный результат для массы случайных явлений оказывается уже закономерным. В значительной мере данное обстоятельство — причина широкого распространения вероятностных методов исследования в биологии и медицине.

Статистика возникла существенно раньше теории вероятностей. Еще в глубокой древности проводились переписи населения и велись земельные кадастры. Эти операции были связаны с наблюдениями и вычислениями. На протяжении веков статистика искала свой математический аппарат и нашла его в теории вероятностей. *В результате возник такой раздел математики, как математическая статистика, в котором устанавливаются закономерности случайных явлений на основании обработки статистических данных — результатов наблюдений и измерений.*

1.2. ВЕРОЯТНОСТЬ СЛУЧАЙНОГО СОБЫТИЯ

Случайное событие — это всякое явление (факт), которое в результате опыта (испытания) может произойти или не произойти. Случайные события часто обозначаются буквами $A, B, C \dots$ и т. д.

Основной количественной характеристикой случайного события является его вероятность. Пусть A — какое-то случайное событие. *Вероятность случайного события — это математическая величина, которая определяет возможность его появления.* Она обозначается $P(A)$. Рассмотрим два основных метода определения этой величины.

Классическое определение вероятности случайного события обычно базируется на результатах анализа умозрительных опытов (испытаний), суть которых определяется условием поставленной задачи. При этом вероятность случайного события $P(A)$ равна:

$$P(A) = \frac{m}{n}, \quad (1)$$

где m — число случаев, благоприятствующих появлению события A ; n — общее число равновозможных случаев.

Пример. Лабораторная крыса помещена в лабиринт и должна выбрать один из пяти возможных путей, так как лишь один из них ведет к поощрению в виде пищи. В предположении равновозможности выбора пути определите вероятность выбора пути, ведущего к пище.

Решение. По условию задачи, из пяти равновозможных случаев ($n = 5$) событию A — «крыса находит пищу» — благоприятствует один из них, т. е. $m = 1$. Тогда

$$P(A) = P(\text{крыса находит пищу}) = \frac{m}{n} = \frac{1}{5} = 0,2 = 20 \, \%.$$

Перечислим свойства вероятности, следующие из ее классического определения:

1. Вероятность случайного события — величина безразмерная.
2. Вероятность случайного события всегда положительна и меньше единицы, т. е. $0 < P(A) < 1$.

3. Вероятность достоверного события, т. е. события которое в результате опыта обязательно произойдет ($m = n$), равна единице.

4. Вероятность невозможного события ($m = 0$) равна нулю.

5. Вероятность любого события — величина не отрицательная и не превышающая единицу: $0 \leq P(A) \leq 1$.

Статистическое определение вероятности случайного события применяется тогда, когда невозможно использовать классическое определение (1). Это часто имеет место в биологии и медицине. В таком случае вероятность $P(A)$ определяют путем обобщения результатов реально проведенных серий испытаний (опытов).

Введем понятие *относительной частоты появления случайного события*. Пусть была проведена серия испытаний, состоящая из N опытов (число N может быть выбрано заранее); интересующее нас событие A произошло в M из них ($M < N$).

Отношение числа опытов M , в которых это событие произошло, к общему числу проведенных опытов N называют *относительной частотой появления случайного события A в данной серии опытов* — $P^*(A)$:

$$P^*(A) = \frac{M}{N} \quad (2)$$

Именно эту величину используют для приближенной оценки статистической вероятности:

$$P(A) \approx P^*(A) = \frac{M}{N}. \quad (3)$$

Чем больше N , тем точнее оценка, тем ближе значения $P^*(A)$ к $P(A)$. Точное значение статистической вероятности события определяется пределом этого отношения при $N \rightarrow \infty$:

$$P(A) = \lim_{N \rightarrow \infty} \left(\frac{M}{N} \right) \quad (3a)$$

Например, в опытах по бросанию монеты относительная частота появления герба при 12 000 бросаний оказалась равной 0,5016, а в серии из 24 000 бросаний — 0,5005. В соответствии с формулой (3)

$$P(\text{появление герба}) = \frac{1}{2} = 0,5 = 50 \, \%.$$

Пример. При врачебном обследовании 500 человек у 5 нашли опухоль в легких (о. л.). Определите относительную частоту и вероятность этого заболевания.

Решение. По условию задачи $M = 5$, $N = 500$, относительная частота $P^*(\text{о. л.}) = M/N = 5/500 = 0,01$. В этой задаче N велико и можно с достаточной точностью считать, что $P(\text{о. л.}) = P^*(\text{о. л.}) = 0,01 = 1 \, \%$.

Перечисленные ранее свойства вероятности случайного события сохраняются и при статистическом определении данной величины.

1.3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ. ВИДЫ СЛУЧАЙНЫХ ВЕЛИЧИН

Величина, которая принимает различные числовые значения под влиянием случайных обстоятельств, называется случайной величиной. Примеры случайных величин: число больных на приеме у врача, точные размеры внутренних органов людей и т. д.

Различают дискретные и непрерывные случайные величины.

Случайная величина называется дискретной, если она принимает только определенные, отделенные друг от друга значения, которые можно установить и перечислить.

Примеры:

1) число студентов в аудитории может быть только целым положительным числом: 0, 1, 2, 3, 4... 20...

2) число событий, происходящих за одинаковые промежутки времени: частота пульса, число вызовов скорой помощи за час, количество операций в месяц с летальным исходом и т. д.

Случайная величина называется непрерывной, если она может принимать любые значения внутри некоторого интервала, который иногда имеет резко выраженные границы, а иногда и нет¹. К непрерывным случайным величинам относятся, например, масса тела и рост взрослых людей, масса и объем мозга, количественное содержание ферментов у здоровых людей, размеры форменных элементов крови, рН крови и т. п.

Если случайная величина зависит от времени, то можно говорить о случайном процессе.

Понятие случайной величины играет определяющую роль в современной теории вероятностей, разработавшей специальные приемы перехода от случайных событий к случайным величинам.

1.4. ЗАКОН РАСПРЕДЕЛЕНИЯ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Чтобы дать полную характеристику дискретной случайной величины необходимо указать все ее значения и их вероятности.

Соответствие между возможными значениями дискретной случайной величины и их вероятностями называется законом распределения этой величины. Обозначим возможные значения случайной величины X через x_i , а соответствующие им вероятности через p_i ². Тогда закон распределения дискретной случайной величины можно задать тремя способами:

¹ В этом случае считают, что значения некоторой случайной величины X могут лежать в интервале $(-\infty; \infty)$, т. е. на всей числовой оси.

² Обычно случайные величины обозначают большими буквами латинского алфавита, а их возможные значения и вероятности этих значений — малыми.

1. В виде таблицы, которая называется рядом распределения:

X	x_1	x_2	$\dots$	x_i	$\dots$	x_n
$P(X)$	p_1	p_2	$\dots$	p_i	$\dots$	p_n

При этом сумма всех вероятностей p_i равна 1 (условие нормировки):

$$p_1 + p_2 + \dots + p_n = \sum_{i=1}^n P(x_i) = 1. \quad (4)$$

2. Графически — в виде ломаной линии (рис. 1), которую принято называть многоугольником распределения:

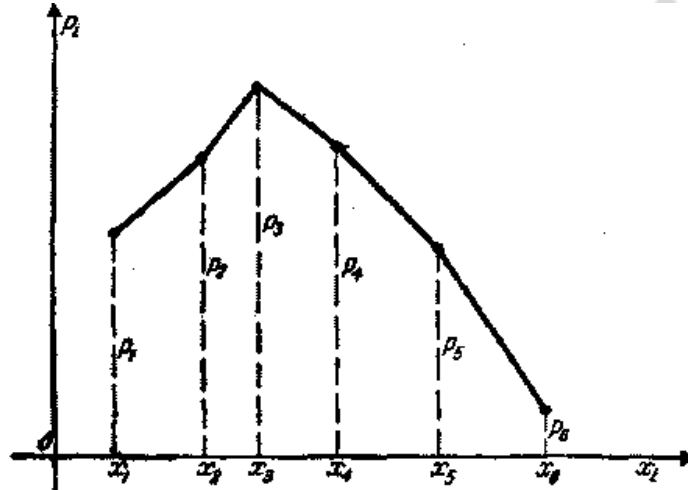


Рис. 1

3. Аналитически — в виде формулы. Например, если вероятность попадания в цель при одном выстреле равна p , то вероятность поражения цели 1 раз при n выстрелах дается формулой $P(n) = n \cdot q^{n-1} \cdot p$, где $q = 1 - p$ — вероятность промаха при одном выстреле.

1.5. ЗАКОН РАСПРЕДЕЛЕНИЯ НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ. ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

Для непрерывных случайных величин невозможно применить закон распределения в формах, приведенных выше, поскольку такая величина имеет бесчисленное («несчетное») множество возможных значений, сплошь заполняющих некоторый интервал. Поэтому составить таблицу, в которой были бы перечислены все ее возможные значения, или построить многоугольник распределения нельзя. Кроме того, вероятность какого-либо ее конкретного значения очень мала (близка к 0)¹. Вместе с тем

¹ Приведем пример, поясняющий этот факт. Пусть случайная величина — уровень осадков, выпавших за год. Она может принимать любые значения из некоторого интервала. Однако вероятность того, что в заданный год этот уровень окажется точно равен 40 см, фактически равна 0.

различные области (интервалы) возможных значений непрерывной случайной величины не равновероятны. Таким образом, и в данном случае действует некий закон распределения, хотя и не в прежнем смысле. Рассмотрим непрерывную случайную величину X , возможные значения которой сплошь заполняют некий интервал (a, b) ¹. Закон распределения вероятностей такой величины должен позволить найти вероятность попадания ее значения в любой заданный интервал (x_1, x_2) , лежащий внутри (a, b) (рис. 2).

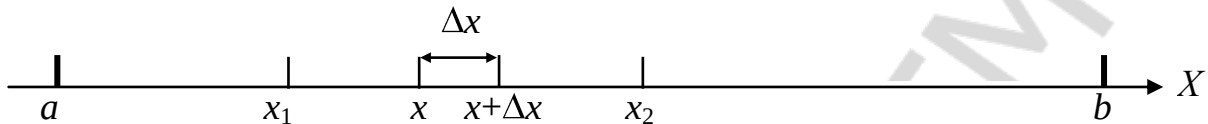


Рис. 2

Эту вероятность обозначают $P(x_1 < X < x_2)$, или $P(x_1 \leq X \leq x_2)$.

В теории вероятностей показано, что ее можно вычислить, введя величину, которая называется плотностью распределения вероятностей случайной величины X , или, короче, плотностью вероятности, плотностью распределения, обозначим ее $f(x)$. Для малого интервала Δx значений X (рис. 2) вероятность того, что случайная величина X примет какое-то значение из этого интервала равна ΔP , тогда

$$\Delta P = f(x) \cdot \Delta x, \text{ а } f(x) = \Delta P / \Delta x. \quad (5)$$

Вероятность попадания значений величины X в конечный интервал (x_1, x_2) (рис. 2) определяется следующей формулой:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x) dx. \quad (6)$$

Графически вероятность $P(x_1 < X < x_2)$ равна площади криволинейной трапеции, ограниченной осью абсцисс, кривой $f(x)$ и прямыми $X = x_1$ и $X = x_2$ (рис. 3). Это следует из геометрического смысла определенного интеграла (6). Кривая $f(x)$ при этом называется *кривой распределения*.

Из (6) и рис. 3 следует, что если известна функция $f(x)$, то, изменяя пределы интегрирования, можно найти вероятность для любых интересующих нас интервалов. Поэтому именно задание функции $f(x)$ полностью определяет закон распределения для непрерывных случайных величин.

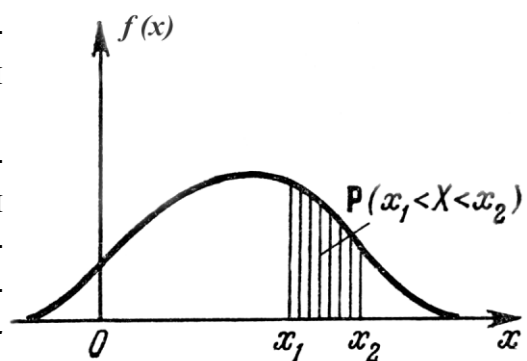


Рис. 3

¹ Иногда рассматривают интервал $(-\infty; +\infty)$.

Для плотности вероятности $f(x)$ должно выполняться условие нормировки в виде

$$\int_a^b f(x)dx = 1, \quad (7)$$

если известно, что все значения X лежат в интервале (a, b) , или в виде

$$\int_{-\infty}^{+\infty} f(x)dx = 1, \quad (8)$$

если границы интервала для значений X точно не определены. Условия нормировки плотности вероятности (7) или (8) являются следствием того, что значения случайной величины X достоверно лежат в пределах (a, b) или $(-\infty, +\infty)$. Из (7) и (8) следует, что *площадь фигуры, ограниченной кривой распределения и осью абсцисс, всегда равна 1*.

1.6. ОСНОВНЫЕ ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

Результаты, изложенные в подразд. 1.4 и 1.5, показывают, что полную характеристику дискретной и непрерывной случайных величин можно получить, зная законы их распределения. Однако во многих практически значимых ситуациях пользуются так называемыми числовыми характеристиками случайных величин. Главное назначение этих характеристик — выразить в сжатой форме наиболее существенные особенности распределения случайных величин. Важно, что данные параметры представляют собой конкретные (постоянные) значения, которые можно оценивать с помощью полученных в опытах данных. Этими оценками занимается «Описательная статистика».

В теории вероятностей и математической статистике используется достаточно много различных характеристик, но мы рассмотрим только наиболее употребляемые, не приводя формулы для их вычисления:

1. *Характеристики положения* — математическое ожидание, мода, медиана.

Они характеризуют положение случайной величины на числовой оси, т. е. указывают некоторое ориентировочное значение случайной величины, около которого группируются все другие ее возможные значения. Среди них важнейшую роль играет математическое ожидание $M(X)$.

Математическое ожидание $M(X)$ случайной величины X является вероятностным аналогом ее среднего арифметического $\bar{X}$: $M(X) = \bar{X}$ или $M(X) \approx \bar{X}$.

Модой $Mo(X)$ дискретной случайной величины называют ее наиболее вероятное значение (рис. 4, а), а непрерывной — значение X , при котором плотность вероятности максимальна (рис. 4, б).

Медианой (Me) случайной величины обычно пользуются только для непрерывных случайных величин, хотя формально ее можно определить и для дискретных X . Медианой $Me(X)$ случайной величины называют такое значение X , которое делит все распределение на две равновероятные части, т. е. вероятности $P(X < Me)$ и $P(X > Me)$ оказываются равными между собой:

$$P(X < Me) = P(X > Me) = \frac{1}{2}.$$

Графически медиана — это значение случайной величины, ордината которой делит площадь, ограниченную кривой распределения, пополам: $S_1 = S_2$ (рис. 4, в).

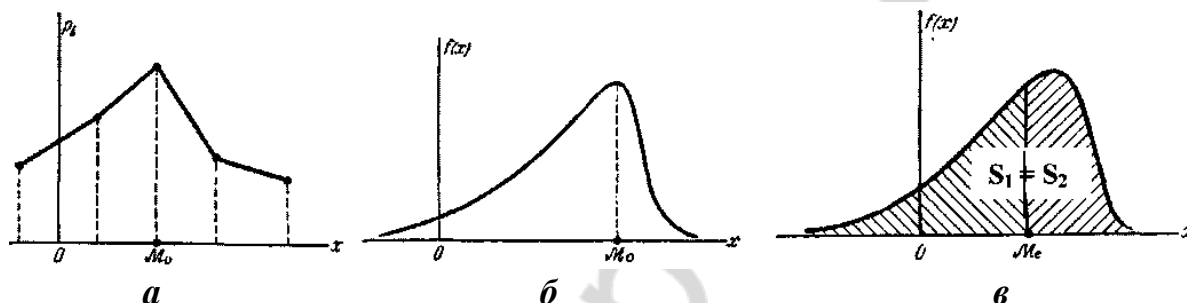


Рис. 4

Если $M(X)$, $Mo(X)$ и $Me(X)$ совпадают, то распределение случайной величины называют симметричным, в противном случае — асимметричным.

2. *Характеристики рассеяния* — это дисперсия и стандартное отклонение.

Дисперсия $D(X)$ случайной величины X характеризует рассеяние, разбросанность значений случайной величины X относительно ее математического ожидания. Само слово «дисперсия» означает «рассеяние».

Дисперсия $D(X)$ имеет размерность квадрата случайной величины. Это весьма неудобно при оценке разброса в физике, биологии, медицине. Поэтому обычно пользуются параметром, размерность которого совпадает с размерностью X . Это *стандартное отклонение* случайной величины X , которое обозначают $\sigma(X)$: $\sigma(X) = \sqrt{D(X)}$.

3. *Характеристики формы* — асимметрия и эксцесс.

Асимметрия As (коэффициент асимметрии) характеризует «скошенность» распределения. Если распределение симметрично относительно математического ожидания, коэффициент асимметрии равен нулю. На рис. 5 показаны два асимметричных распределения. Одно из них (кривая

I) имеет положительную асимметрию ($As > 0$), другое (кривая II) — отрицательную ($As < 0$).

Экссесс Ex (коэффициент эксцесса) используется для характеристики так называемой «крутости», т. е. островершинности или плосковершинности распределения. Для нормального распределения (см. подразд. 1.7) эксцесс равен 0. Кривые, более островершинные по сравнению с нормальной, обладают положительным эксцессом, кривые, более плосковершинные, — отрицательным эксцессом. На рис. 6 представлены: нормальное распределение (кривая I), распределение с положительным эксцессом (кривая II) и распределение с отрицательным эксцессом (кривая III).

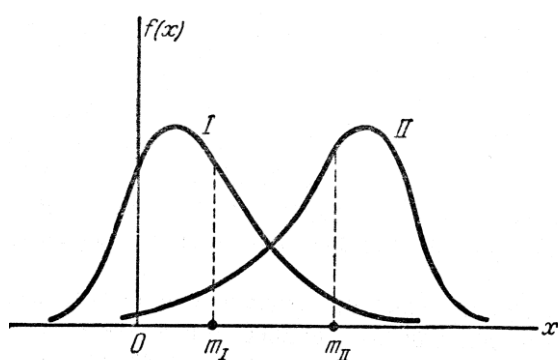


Рис. 5

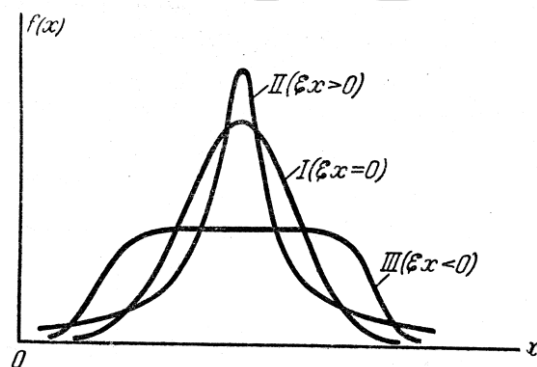


Рис. 6

Итак, математическое ожидание, мода, медиана, дисперсия, стандартное отклонение, асимметрия и эксцесс являются наиболее употребляемыми числовыми характеристиками случайных величин, каждая из которых выражает какое-нибудь характерное свойство их распределения.

1.7. НОРМАЛЬНЫЙ ЗАКОН РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

Нормальный закон распределения (закон Гаусса) играет исключительно важную роль в теории вероятностей. Во-первых, это наиболее часто встречающийся на практике закон распределения непрерывных случайных величин. Во-вторых, он является предельным законом в том смысле, что к нему при определенных условиях приближаются другие законы распределения.

Нормальный закон распределения характеризуется следующей формулой для плотности вероятности:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{[x-M(x)]^2}{2\sigma^2}}, \quad (9)$$

где x — текущие значения случайной величины X ; $M(X)$ и σ — ее математическое ожидание и стандартное отклонение. Из (9) видно, что если слу-

чайная величина распределена по нормальному закону, то достаточно знать только два числовых параметра — $M(X)$ и σ — чтобы полностью знать закон ее распределения.

График функции (9) называется нормальной кривой распределения (кривой Гаусса). Он имеет симметричный вид относительно прямой $x = M(X)$ (рис. 7).

Величина $M(X)$ называется также центром рассеяния. Стандартное отклонение σ характеризует ширину кривой распределения.

При изменении значения $M(X)$ в (9)

нормальная кривая не меняется по форме, но сдвигается вдоль оси абсцисс. С возрастанием σ макси-

мальное значение $f_{\max}(x) = \frac{1}{\sigma\sqrt{2\pi}}$

убывает, а сама кривая, становясь более пологой, растягивается вдоль оси абсцисс; при уменьшении σ кривая вытягивается вверх, одновременно сжимаясь с боков. Вид кривой распределения при разных значениях σ : ($\sigma_3 < \sigma_2 < \sigma_1$) показан на рис. 8.

Естественно, что при любых значениях $M(X)$ и σ площадь, ограниченная нормальной кривой и осью X , остается равной 1 (условие нормировки):

$$\int_a^b f(x)dx = 1, \text{ или } \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Нормальное распределение симметрично, поэтому среднее, мода и медиана равны друг другу: $M(X) = Mo(X) = Me(X)$, асимметрия $As = 0$, эксцесс $Ex = 0$.

Вероятность попадания значений случайной величины X в интервал (x_1, x_2) , т. е. $P(x_1 < X < x_2)$, равна:

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{[x-M(X)]^2}{2\sigma^2}} dx. \quad (10)$$

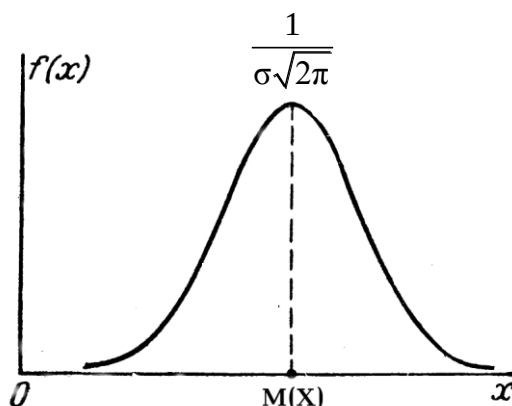


Рис. 7

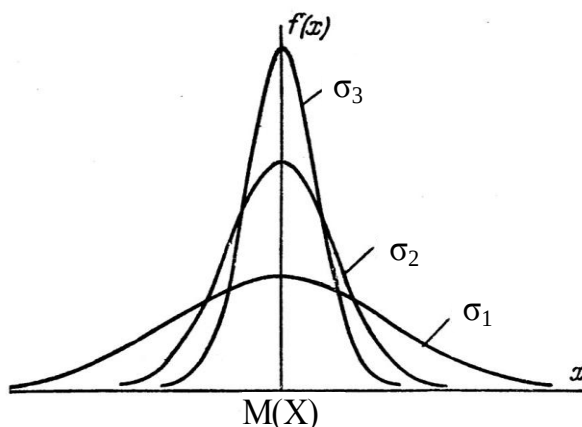


Рис. 8

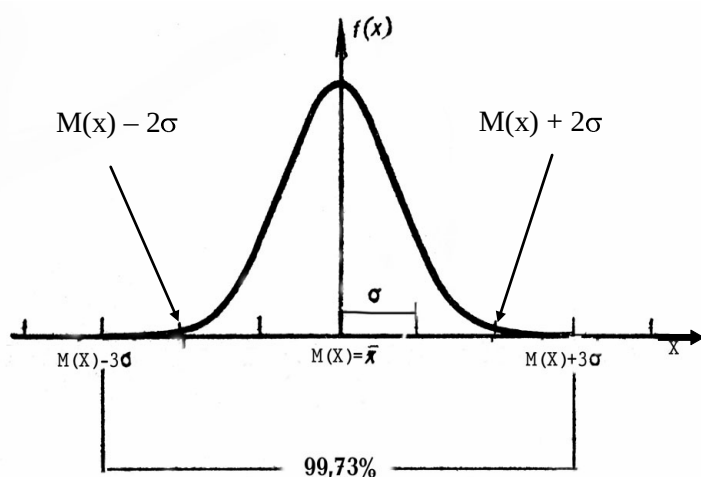


Рис. 9

На практике часто приходится вычислять вероятности попадания значений нормально распределенной случайной величины на участки, симметричные относительно $M(X)$. В частности, рассмотрим следующую, важную в прикладном отношении задачу. Отложим от $M(X)$ вправо и влево отрезки, равные σ , 2σ и 3σ (рис. 9), и проанализируем результат

вычисления вероятности попадания значений X в соответствующие интервалы:

$$P(M(X) - \sigma < X < M(X) + \sigma) = 0,6827 = 68,27 \%. \quad (11)$$

$$P(M(X) - 2\sigma < X < M(X) + 2\sigma) = 0,9545 = 95,45 \%. \quad (12)$$

$$P(M(X) - 3\sigma < X < M(X) + 3\sigma) = 0,9973 = 99,73 \%. \quad (13)$$

Из (13) следует: практически достоверно, что значения нормально распределенной случайной величины X с параметрами $M(X)$ и σ лежат в интервале $M(X) \pm 3\sigma$. Иначе говоря, зная $M(X) = \bar{X}$ и σ , можно указать интервал, в который с вероятностью $P = 99,73 \%$ попадают значения данной случайной величины. Такой способ оценки диапазона возможных значений X известен как «правило трех сигм».

Пример. Известно, что для здорового человека рН крови является нормально распределенной величиной со средним значением (математическим ожиданием) 7,4 и стандартным отклонением 0,2. Определите диапазон значений этого параметра.

Решение. Для ответа на этот вопрос воспользуемся «правилом трех сигм». С вероятностью, равной 99,73 %, можно утверждать, что диапазон значений рН для здорового человека составляет 6,8–8.

Глава 2

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

2.1. ПРЕДМЕТ И ЗАДАЧИ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ. ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТЬ

Предмет математической статистики — это разработка методов получения, описания и анализа статистических данных, определенных в результате исследования массовых случайных явлений. Статистические данные часто можно рассматривать как совокупность экспериментальных результатов, которые представляют собой набор возможных значений случайных величин (роста, массы тела, длительности пребывания больного на койке, содержания сахара в крови и т. д.).

Фундаментальными понятиями математической статистики являются генеральная и выборочная совокупности (выборка). Существуют разные подходы к пониманию смысла этих величин. Мы определяем их так. *Генеральная совокупность* — это множество подлежащих статистическому изучению однородных объектов, которые характеризуются определенными качественными или количественными признаками. Например, конечная и реально существующая генеральная совокупность — конкретно выбранная популяция: все жители Беларуси в фиксированный момент времени или только все мужчины, или женщины, или дети. Следующий пример: бесконечная и реально существующая генеральная совокупность — множество чисел, лежащих между 0 и 1.

Чтобы изучить генеральную совокупность по какому-либо из ее количественных признаков X (острота зрения, показатели анализа крови и т. д.), нужно определить закон распределения данного признака и основные характеристики этого распределения, например, математическое ожидание и дисперсию. Для этого следовало бы изучить все ее объекты и затем обработать полученный массив данных методами теории вероятностей. Однако на практике провести сплошное обследование объектов генеральной совокупности часто физически невозможно и экономически невыгодно. Поэтому обычно исследуется только часть объектов, так называемая выборка.

Совокупность «n» объектов, отобранных из интересующей нас генеральной совокупности для конкретного статистического исследования, называется выборочной совокупностью, или выборкой.

Исследование выборки дает некоторое приближенное, оценочное значение интересующего нас параметра, принимающего различные значения для разных выборок. Поэтому главная цель выборочного метода, основного в математической статистике, — по вычисленной характеристике выборки как можно точнее определить соответствующую характе-

ристику генеральной совокупности. Это возможно лишь в том случае, когда отобранная для работы часть объектов репрезентативна целому, т. е. типична, обладает теми же основными чертами, что и все целое. Иначе говоря, выборка должна быть представительной, т. е. по возможности полнее «представлять» свою генеральную совокупность. Это одно из важнейших требований, предъявляемых к выборке, несоблюдение которого ведет к грубым ошибкам и обесценивает результаты исследования. Например, если при изучении заболеваемости населения республики (генеральная совокупность) ишемической болезнью сердца в качестве выборки будет взята группа студентов, то результаты окажутся ошибочными, поскольку свойства выборки не будут соответствовать свойствам генеральной совокупности, то же будет, когда в качестве выборки взяты только пациенты кардиологического диспансера. Репрезентативность выборки обеспечивается ее достаточным объемом и определенными правилами ее формирования, которые в данном издании не рассматриваются.

Из многочисленных **задач**, решаемых математической статистикой, выделим следующие:

1. Определение статистических характеристик выборки (методы описательной статистики).
2. Определение параметров генеральной совокупности по данным выборки: точечные оценки и доверительные интервалы для параметров распределения.
3. Проверка статистических гипотез.
4. Исследование статистической связи между двумя признаками выборочной совокупности (элементы корреляционного анализа).

В данной главе излагаются общие подходы к решению этих задач. Конкретные примеры разобраны, главным образом, в гл. 3.

2.2. СТАТИСТИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ ВЫБОРКИ

Обычно необходимо знать распределение признака X в генеральной совокупности, но реально исследуется лишь некоторая выборка из нее.

В серии экспериментов, проводимых с выборкой, величина X принимает определенные значения. Значения, записанные для всех элементов выборки в том порядке, в котором они были получены в опытах, представляют собой *простой статистический ряд*: $x_1, x_2, x_3 \dots x_n$. Каждое значение X в полученном числовом ряду называют *вариантой*. Полученные данные и подлежат статистической обработке, статистическому анализу.

Первый шаг при обработке этого материала — наведение в нем определенного порядка, ведущего к получению статистического распределения выборки. Здесь возможны два основных способа: создание вариационного или интервального ряда.

Рассмотрим *вариационный ряд*. Пусть некоторая выборка исследуется по количественному признаку X , который представляет собой дискретную случайную величину. В имеющемся у нас простом статистическом ряду варианта x_1 встречается (повторяется) m_1 раз, x_2 — m_2 раза, ... x_k —

m_k раз, при этом $\sum_{i=1}^k m_i = n$, т. е. равна объему выборки. Далее по данным

простого статистического ряда строится статистическое распределение (в медицинской литературе — *вариационный ряд*), которое удобно представить в виде таблицы, включающей в себя:

1) различные по значению варианты x_i , расположенные в определенной, ранжированной¹, заранее выбранной последовательности (обычно в порядке возрастания);

2) m_i — частоты вариант, т. е. числа наблюдений (повторений) варианты x_i в простом статистическом ряду;

3) $p_i = m_i / n$ — относительные частоты вариант, т. е. отношения частот m_i к объему выборки n ; они являются выборочными (эмпирическими) оценками вероятностей появления значений x_i (см. 1.2).

Итак, для дискретной величины X вариационный ряд — статистическое распределение выборки — имеет следующий вид (табл. 1).

Таблица 1

Варианта x_i ($x_1 < x_2 < x_3 \dots < x_k$)	x_1	x_2	x_3	...	x_k	Контроль
Частота m_i	m_1	m_2	m_3	...	m_k	$\sum_{i=1}^k m_i = n$
Относительная частота $p_i^* = \frac{m_i}{n}$	$\frac{m_1}{n}$	$\frac{m_2}{n}$	$\frac{m_3}{n}$	...	$\frac{m_k}{n}$	$\sum_{i=1}^k \frac{m_i}{n} = 1$

Напомним, что под распределением дискретной случайной величины в теории вероятностей понимается соответствие между возможными значениями случайной величины и их вероятностями; в математической статистике — соответствие между наблюдаемыми вариантами x_i и их частотами или относительными частотами.

Интервальный ряд удобен тогда, когда количественный признак X , характеризующий выборку, непрерывен, т. е. может принимать любые значения в некотором интервале. В этом случае статистическое распределение выборки (интервальный ряд) строится следующим образом. Область изменения признака ($x_{\max} - x_{\min}$) разбивают на несколько интервалов обычно равной ширины. Число интервалов k , как правило, не менее

¹ В математической статистике ранжированным рядом часто называется последовательность всех полученных в эксперименте вариант, записанных в порядке возрастания.

5 и не более 25 и приближенно определяется следующими эмпирическими формулами:

$$k \approx \sqrt{n}, \text{ или } k \approx 1 + 3,32 \lg n, k \approx 5 \lg n \quad (14)$$

где n — объем выборки.

Если ширина интервалов одинакова, то она равна:

$$\Delta x = h = \frac{x_{\max} - x_{\min}}{k} \quad (15)$$

Затем вычисляют границы интервалов: $x_{\min} = x_0$, $x_1 = x_0 + h$, $x_2 = x_1 + h$, $x_3 = x_2 + h, \dots, x_{\max} = x_k$. Поскольку некоторые варианты могут являться границей двух соседних интервалов, то, во избежание недоразумений, придерживаются следующего правила: к интервалу (a, b) относят варианты, удовлетворяющие неравенству $a \leq x < b$.

Затем для каждого интервала подсчитывают частоты m_i и (или) относительные частоты $p_i = m_i/n$ попадания вариантов в данный интервал (эмпирические оценки вероятности попадания значений X в выбранный интервал). Нередко используют также плотность относительной частоты:

$$\frac{m_i}{n \Delta x} = \frac{m_i}{n h}.$$

Данную величину можно считать выборочной (эмпирической) оценкой плотности вероятности.

Рассмотренное выборочное распределение непрерывной случайной величины X — интервальный ряд — обычно представляется в виде таблицы, имеющей следующий вид (табл. 2).

Таблица 2

Интервал	$x_0 - x_1$	$x_1 - x_2$	...	$x_{k-1} - x_k$	Контроль
Частота m_i	m_1	m_2	...	m_k	$\sum_{i=1}^k m_i = n$
Относительная частота $p_i^* = m_i/n$	m_1/n	m_2/n	...	m_k/n	$\sum_{i=1}^k \frac{m_i}{n} = 1$
Плотность относительной частоты $p_i^*/\Delta x = m_i/n \Delta x$	$m_1/n \Delta x$	$m_2/n \Delta x$	...	$m_k/n \Delta x$	$\sum_{i=1}^k \frac{m_i}{n \cdot \Delta x} = \frac{1}{\Delta x}$

Обобщим изложенный выше материал:

1. Если выборка исследуется по количественному признаку X , который представляет собой дискретную случайную величину, то статистическим распределением выборки является вариационный статистический ряд — полученные разные значения признака, записанные в упорядоченном виде с указанием их частот и относительных частот.

2. Если выборка исследуется по количественному признаку X , который представляет собой непрерывную случайную величину, то статисти-

ческим распределением выборки является интервальный статистический ряд. Он включает в себя интервалы вариант, частоты попадания вариант в эти интервалы, относительные частоты, при необходимости — плотности относительных частот для этих интервалов.

2.3. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКИХ РАСПРЕДЕЛЕНИЙ ВЫБОРОК

Для получения наглядного представления о распределении выборок строят соответствующие графики, в частности, полигон частот или гистограмму распределения.

Вариационный ряд часто изображают графически в виде **полигона частот** или **полигона относительных частот**.

Для построения полигона частот на оси абсцисс откладывают варианты x_i , а на оси ординат — соответствующие им частоты m_i . Точки $(x_i; m_i)$ соединяют отрезками прямых. *Полигоном частот называют ломаную линию, отрезки которой соединяют точки (x_1, m_1) ; (x_2, m_2) ... (x_k, m_k) .*

Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки $(x_1, \frac{m_1}{n})$; $(x_2, \frac{m_2}{n})$; $(x_k, \frac{m_k}{n})$.

На рис. 10 показан полигон относительных частот, построенный по данным выборки для некоторой случайной величины.

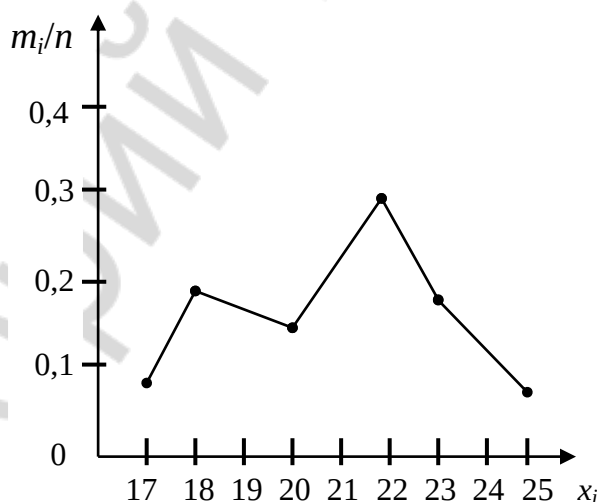


Рис. 10

Для непрерывной случайной величины обычно строят **гистограммы**.

Гистограммой называют диаграмму, состоящую из вертикальных прямоугольников, основаниями которых являются интервалы длиной $\Delta x = h$, а высоты равны m_i (частоте), m_i/n^1 (относительной частоте), m_i/n %, отношению $\frac{m_i}{\Delta x}$ или $\frac{m_i}{\Delta x n}$ для соответствующих интервалов.

В случае, когда строят прямоугольники высотой $\frac{m_i}{\Delta x}$, площадь каждого из них равна количеству вариант в i -м интервале, т. е. площадь ги-

¹ Эти варианты позволяют сравнить гистограммы, построенные на одних и тех же интервалах, но для различных выборок из той же генеральной совокупности.

гистограммы равна сумме частот для всех интервалов, иначе говоря, равна объему выборки. Такую гистограмму называют *гистограммой частот*.

Если строят прямоугольники высотой $\frac{m_i}{\Delta x \cdot n}$, то площадь каждого i -го

прямоугольника является оценкой вероятности попадания значений x в выбранный интервал. В этом случае площадь гистограммы равна единице, а гистограмма называется *гистограммой относительных частот* (рис. 11, здесь анализируемый показатель — масса тела новорожденного).

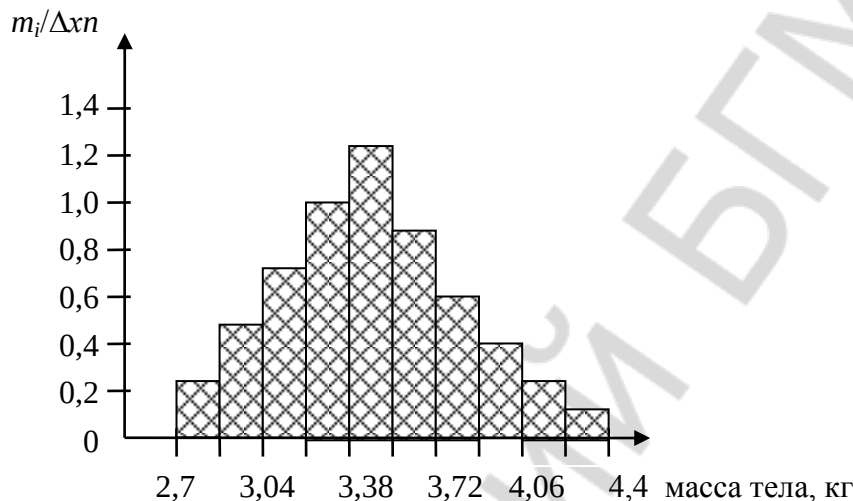


Рис. 11

Важно, что гистограммы можно использовать для оценки закона распределения признака в генеральной совокупности (в популяции). Соединяя средние точки верхних оснований прямоугольников гистограммы относительных частот плавной линией, можно по данным выборки получить примерный вид графика зависимости плотности вероятности f от x .

2.4. МЕТОДЫ ОПИСАТЕЛЬНОЙ СТАТИСТИКИ

Это методы описания выборок, исследуемых по количественному признаку X , с помощью их различных числовых характеристик.

Преимущество данных методов заключается в следующем. Несколько простых и достаточно информативных статистических показателей, если они известны, во-первых, избавляют нас от просмотра сотен, а порой и тысяч значений вариантов, а во-вторых, позволяют получить более или менее точную оценку характеристик распределения признака в генеральной совокупности.

Описывающие выборку показатели разбиваются на несколько групп; в своем большинстве они имеют аналоги в виде числовых характеристик случайных величин в теории вероятностей.

Показатели положения описывают положение вариант выборки на числовой оси. Сюда относят:

- а) минимальную и максимальную варианты;
- б) выборочное среднее арифметическое значение (выборочное среднее), выборочные моду и медиану. Они определяют «центральную» точку распределения выборки — наиболее значимую для поставленной задачи варианту.

Выборочным средним называется величина

$$\bar{x}_B = \frac{\sum_{i=1}^n x_i}{n}, \quad (16)$$

где x_i — i -я варианта, полученная в опыте с i -м элементом выборки; n — объем выборки.

Выборочное среднее является той точкой, сумма отклонений значений X от которой равна нулю. Это единственная точка, которая обладает данным свойством, оно выделяет ее среди всех других.

Выборочная мода Mo_v — варианта, которая чаще всего встречается в исследуемой выборке, т. е. имеет наибольшую частоту.

Пример 1. На рис. 12 приведено предполагаемое распределение по возрасту заболевших дифтерией (на 10 тыс. населения соответствующего возраста), которое явно не соответствует нормальному. Очевидно, что знание среднего возраста заболевших ($\bar{x}_B \approx 7,8$ года = μ)

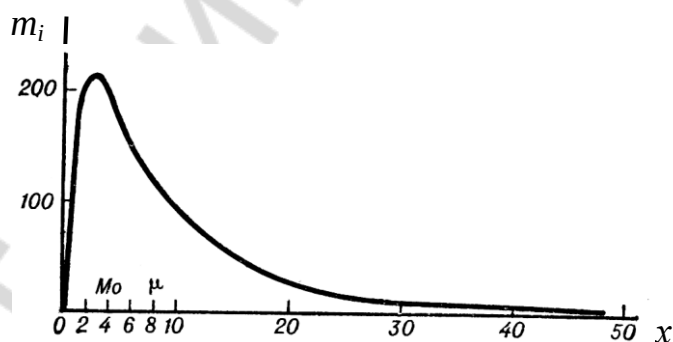


Рис. 12

в этом случае менее важно, чем знание возраста, в котором чаще всего возникает заболевание и который представляет собой моду ($Mo_v \approx 4$ года). Именно этот показатель указывает, где должны быть сосредоточены главные профилактические меры: в школах или дошкольных учреждениях.

Если выборочное распределение имеет несколько мод, то говорят, что оно мультимодально. Это служит индикатором того, что выборка не является однородной, и данные, возможно, порождены несколькими «наложенными» распределениями.

Выборочная медиана Me_v — варианта, которая делит ранжированный статистический ряд (см. сноску на стр. 17) на две равные части по числу попадающих в них вариант.

Пример 2. Дан статистический ряд: 1; 2; 3; 3; 4; 4; 5; 5; 6; 8; 9; $n = 11$. Варианта, разделяющая этот ряд на две равные по количеству вариант части, занимает в ряду 6 место и равна 4, т. е. $Me_v = 4$.

Показатели разброса описывают степень разброса данных относительно своего центра. Здесь обычно используются:

а) *стандартное отклонение* S и *выборочная дисперсия* $D_s = S^2$ ¹, характеризующие рассеяние вариант вокруг их среднего выборочного значения $\bar{x}_B$:

$$S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x}_B)^2}{n-1}}. \quad (17)$$

Именно эти параметры часто используются как меры изменчивости некоторого исследуемого показателя (случайной величины X). Чем больше D_B и S , тем сильнее разбросаны значения X относительно среднего;

б) *размах выборки* — разность между максимальной и минимальной вариантами: $x_{\max} - x_{\min}$;

в) *коэффициент вариации*:

$$v = \frac{S}{\bar{x}_B} \cdot 100 \%. \quad (18)$$

Применяется для сравнения величин рассеяния двух вариационных рядов: тот из них имеет большее рассеяние, у которого коэффициент вариации больше.

Отметим так же выборочный эксцесс (Ex_B) и выборочную асимметрию As_B (их смысл — см. подразд. 1.6).

Для определения **закона распределения** исследуемой величины в генеральной совокупности, который нужно знать для проведения последующего анализа, можно использовать *гистограммы* и *полигон частот*. О них шла речь в предыдущем подразделе. О законе распределения также можно судить по выборочным числовым характеристикам случайной величины.

Заметим, что большинство методов статистического анализа данных разработано для случайных величин, распределенных по нормальному закону. Распределение исследуемой величины в генеральной совокупности можно рассматривать как близкое к нормальному, если:

1. Выборочные $\bar{x}_B$, Me_B , Mo_B равны или незначительно отличаются друг от друга.
2. Минимальное и максимальное значения X ($x_{\max}$ и $x_{\min}$) примерно равноудалены от $\bar{x}_B$.
3. Выборочные Ex_B и As_B близки к нулю.

¹ Точнее S^2 называется «исправленная выборочная дисперсия».

Можно так же проверить выполнение следующих условий:

1. Примерно 99,7 % отклонений значений исследуемого показателя X от среднего по модулю меньше $3S$.

2. Примерно 95,5 % отклонений значений исследуемого показателя X от среднего по модулю меньше $2S$.

3. Примерно 68,3 % отклонений значений исследуемого показателя X от среднего по модулю меньше S .

Такой подход является следствием расчета вероятностей попадания значений нормально распределенного исследуемого показателя в интервалы: $M(x) \pm \sigma$, $M(x) \pm 2\sigma$, $M(x) \pm 3\sigma$ (см. подразд. 1.7).

2.5. ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО ЕЕ ВЫБОРКЕ. ТОЧЕЧНАЯ И ИНТЕРВАЛЬНАЯ ОЦЕНКИ

Напомним, что главная цель любого статистического исследования — установить закон распределения и получить значения характеристик изучаемого признака генеральной совокупности путем анализа выборки. Иначе говоря, надо определить генеральную среднюю $\bar{x}_g = M(X)$, генеральную дисперсию $D_g(X)$, стандартное отклонение σ_g , генеральную моду Mo_g , медиану Me_g и другие характеристики генеральной совокупности путем статистического исследования выборки.

Точечная оценка характеристик генеральной совокупности — наиболее простой, но не очень достоверный способ. При данном способе в качестве оценок характеристик генеральной совокупности используются соответствующие числовые характеристики выборки. Например, в качестве генерального среднего используется выборочное среднее, в качестве генеральной дисперсии — выборочная дисперсия и т. д. Такие оценки и называются точечными. Их недостаток состоит в том, что не ясно, насколько сильно они отличаются от истинных значений параметров генеральной совокупности. Ошибка может быть особенно большой в случае малых выборок.

Интервальная оценка параметров генеральной совокупности более достоверна. В этом случае определяется интервал, в который с заданной вероятностью попадает истинное значение исследуемого признака. Такой интервал называется *доверительным интервалом*, а вероятность того, что истинное значение оцениваемой величины находится внутри этого интервала — *доверительной вероятностью*, или *надежностью*. В медицинской литературе для этой величины используется термин «*вероятность безошибочного прогноза*». Обозначим ее γ . Значения γ задаются заранее (обычно в медико-биологических исследованиях выбирают значения $\gamma = 0,95 = 95\%$ или $\gamma = 0,99 = 99\%$), после чего находят соответствующий доверительный интервал. Иногда вместо доверительной веро-

ятности используется величина $\alpha = 1 - \gamma$, которая называется уровнем значимости.

Для построения надежных интервальных оценок необходимо знать закон, по которому оцениваемый случайный признак распределен в генеральной совокупности.

Рассмотрим, вначале для малых выборок ($n < 30$), как строится интервальная оценка генеральной средней $\bar{x}_r = M(X)$ признака, который в генеральной совокупности распределен по нормальному закону. В этом случае интервальной оценкой (с доверительной вероятностью γ) генеральной средней (математического ожидания, $\bar{x}_r = M(X)$) количественного признака X по выборочной средней $\bar{x}_b$ при неизвестном σ_r является доверительный интервал

$$\bar{x}_b - \delta < M(X) < \bar{x}_b + \delta, \quad (19)$$

или, в другой форме записи:

$$M(X) = \bar{x}_b \pm \delta, \quad (20)$$

где $\delta = t_{\gamma,n}(S/\sqrt{n})$ — полуширина доверительного интервала — предельная ошибка выборки, характеризующая точность оценки; n — объем выборки; S — выборочное стандартное отклонение; $S/\sqrt{n}$ — стандартная ошибка выборочного среднего (в медицинской и биологической литературе эта величина иногда обозначается буквой m и называется ошибкой репрезентативности), $t_{\gamma,n}$ — коэффициент Стьюдента (его значения определяются либо по соответствующим таблицам, либо содержатся в программных статистических пакетах обработки данных).

Анализ формулы (19) показывает, что:

- а) чем больше доверительная вероятность γ , тем больше коэффициент $t_{\gamma,n}$ и шире доверительный интервал;
- б) чем больше объем выборки n , тем уже доверительный интервал.

При большой выборке ($n > 30$) полуширину доверительного интервала δ определяют по соотношениям:

$$\delta \approx 1,96S/\sqrt{n} \text{ при } \gamma = 95 \% \text{ или } \delta \approx 2,6S/\sqrt{n} \text{ при } \gamma = 99 \%. \quad (21)$$

Рассматривая интервальную оценку $M(X)$, обратим внимание на следующие обстоятельства. Иногда экспериментаторы приводят результат в виде:

$$M(X) = \bar{x}_b \pm S/\sqrt{n} = \bar{x}_b \pm m.$$

По существу, такая запись содержит указание доверительного интервала при $t_{\gamma,n} = 1$. Рассмотрим, на примере, к чему это может привести.

Пример. Произведено 31 измерение какой-то величины $n = 31$. Необходимо определить доверительный интервал с доверительной вероятностью 90 % для математического ожидания $M(X)$ этой величины.

Обработка полученных данных дает: $\bar{x}_B = 58,29$ ед., $S = 5,58$ ед., $t_{0,9;31} = 1,7$. Тогда $M(X) = \bar{x}_B \pm t_{\gamma,n} \cdot (S/\sqrt{n}) = (58,29 \pm 1,7)$ ед. Если указать результат в виде $M(X) = \bar{x}_B \pm S/\sqrt{n}$ ($t_{\gamma,n} = 1$), то в нашем примере $M(X) = 58,29 \pm 1$. По таблицам значений коэффициента Стьюдента легко определить, что при $t_{\gamma,31} = 1$ доверительная вероятность γ равна лишь 68 %. Это резко снижает доверие к полученному результату (с 90 % при правильном расчете до 68 %) несмотря на сужение доверительного интервала.

Из формул (19), (21) понятно, как при заданной доверительной вероятности и объему выборки получить оценку $\bar{x}_r = M(X)$.

Поставим обратную, практически значимую задачу. По заданной точности оценки δ , т. е. по заданной полуширине доверительного интервала, определим необходимый объем выборки, обеспечивающий нужное δ . Эта задача решается особенно просто в случае больших выборок ($n > 30$). Здесь, например, при доверительной вероятности 95 % $\delta = 1,96S/\sqrt{n}$. Тогда из (21) следует, что необходимый объем выборки равен: $n \geq (1,96)^2 S^2/\delta^2$.

Определим доверительный интервал для D_r и σ_r , предполагая, что случайная величина в генеральной совокупности опять же распределена по нормальному закону, а ее математическое ожидание неизвестно. Тогда при заданной доверительной вероятности γ или уровне значимости α , он определяется следующими соотношениями:

а) для дисперсии:

$$\frac{(n-1)S^2}{\chi_{n-1,(1-\gamma)/2}^2} < D_r < \frac{(n-1)S^2}{\chi_{n-1,(1+\gamma)/2}^2}; \quad (22)$$

б) для стандартного отклонения:

$$S \cdot \sqrt{\frac{(n-1)}{\chi_{n-1,(1-\gamma)/2}^2}} < \sigma_r < S \cdot \sqrt{\frac{(n-1)}{\chi_{n-1,(1+\gamma)/2}^2}}. \quad (23)$$

В этих формулах значения χ^2 (хи-квадрат) определяются либо по соответствующим таблицам, либо с помощью встроенных функций программ обработки статистических данных.

Если вместо доверительной вероятности γ использовать уровень значимости α и учесть, что $\alpha = 1 - \gamma$, то формулы для расчета доверительных интервалов принимают следующий вид:

а) для дисперсии:
$$\frac{(n-1)S^2}{\chi_{n-1,\alpha/2}^2} < D_r < \frac{(n-1)S^2}{\chi_{n-1,(1-\alpha/2)}^2}$$

б) для стандартного отклонения:
$$S \cdot \sqrt{\frac{(n-1)}{\chi_{n-1,\alpha/2}^2}} < \sigma_r < S \cdot \sqrt{\frac{(n-1)}{\chi_{n-1,(1+\alpha/2)}^2}}.$$

Подобные интервальные оценки с заданной надежностью даются и в тех случаях, когда рассматриваемый случайный признак распределен в генеральной совокупности не по нормальному, а по другим законам. Соответствующие формулы для вычисления будут другими.

2.6. ПОНЯТИЕ О СТАТИСТИЧЕСКИХ ГИПОТЕЗАХ И КРИТЕРИЯХ ПРОВЕРКИ ГИПОТЕЗ

Во многих случаях требуется на основе экспериментальных данных решить, справедливо ли некоторое утверждение. Например, верно ли, что два набора данных (2 выборки) происходят из одного источника (из одной генеральной совокупности), или что A лучший стрелок, чем B , или что данное лекарство лучше другого при лечении определенного заболевания? При ответе на подобные вопросы, во-первых, хотелось бы принять наиболее обоснованное решение, во-вторых, оценить вероятность ошибки этого решения.

Рассмотрение таких задач в строгой математической постановке приводит к понятию **статистической гипотезы**.

В обычном языке понятие «гипотеза» означает предположение. В математической статистике *гипотеза* — это: а) *предположение о виде неизвестного закона распределения исследуемой экспериментально случайной величины (непараметрическая гипотеза)*; б) *предположение о значениях характеристик (параметров) известного распределения (параметрическая гипотеза)*.

Примеры статистических гипотез. Делаются предположения:

1. Данный признак в генеральной совокупности распределен по нормальному закону (непараметрическая гипотеза).
2. Дисперсии двух совокупностей, распределенных по нормальному закону, равны между собой (параметрическая гипотеза).

Эти предположения подлежат проверке.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей. Если выдвинутая гипотеза отвергнута, то имеет место противоречащая гипотеза.

Выдвинутую гипотезу H_0 называют нулевой (основной). Конкурирующую (альтернативную) гипотезу, которая несовместна с нулевой, обозначают H_1 . Например, если H_0 состоит в предположении, что математическое ожидание $M(X)$ нормального распределения равно 2, то конкурирующая гипотеза, в частности, может состоять в предположении, что $M(X) \neq 2$. Коротко это записывают так:

$$H_0: M(X) = 2;$$

$$H_1: M(X) \neq 2.$$

Заключение о справедливости нулевой или альтернативной гипотезы всегда делается на основании анализа выборки определенного объема.

Если для исследуемого явления сформулирована та или иная гипотеза, то надо найти правило, которое позволяло бы по имеющимся статистическим данным (по выборке) принять решение о соответствии либо несоответствии выдвинутой гипотезы этим данными. Это правило называется *статистическим критерием* (иногда просто критерием, статистикой) проверки гипотезы.

Для проверки непараметрических гипотез существуют *критерии согласия*, которые должны подтвердить или опровергнуть правильность выбора закона распределения. В данном случае чаще других используются *критерий χ^2 (хи-квадрат) Пирсона и критерий Колмогорова–Смирнова*. Первый из них более универсален, так как приемлем для случайных величин любого типа (дискретных и непрерывных), второй — только для непрерывных случайных величин. В данном издании эти критерии лишь упоминаются. При необходимости ими можно воспользоваться, работая со специальными статистическими пакетами (Biostat, Statistica и т. д.) или с соответствующей литературой, например [4, 5, 9].

Параметрические гипотезы проверяются с помощью *параметрических критериев значимости*, разработанных, прежде всего, для случайных величин, распределенных по нормальному закону.

Создание критериев потребовало разработки достаточно сложной теории. Рассмотрим лишь основные ее идеи и продемонстрируем на примерах работу соответствующих правил.

Обычно проверяется нулевая гипотеза (H_0). Тогда статистическим критерием проверки H_0 называют случайную величину (статистику) K , точное или приближенное распределение которой известно. В конкретных задачах K имеет и свое конкретное обозначение. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве статистики K используют отношение выборочных дисперсий (критерий F Фишера–Снедекора, иногда просто критерий Фишера):

$$K = F = S_1^2 / S_2^2 \quad (S_1 > S_2).$$

Так как входящие в критерий величины рассчитывают по данным определенных выборок, то вычисленное значение K называют *наблюдаемым значением критерия* $K_{\text{набл.}}$. Например, если по выборкам $S_1^2 = 20$, а $S_2^2 = 5$, то $K_{\text{набл.}} = F_{\text{набл.}} = 20/5 = 4$.

После выбора определенного критерия множество всех его возможных значений разделяют на две области:

1. *Критическая область* — совокупность значений критерия K , при которых нулевую гипотезу (H_0) отвергают.

2. *Область принятия гипотезы (область допустимых значений)* — совокупность значений критерия K , при которых нет оснований в рамках данного критерия отвергнуть гипотезу H_0 .

Таким образом, основной принцип проверки гипотезы H_0 достаточно прост: если вычисленное по выборке значение критерия $K_{\text{набл.}}$ принадлежит критической области, то гипотезу H_0 отвергают; если оно принадлежит области принятия гипотезы, то H_0 можно принять.

Указанные области (интервалы) разделяются *критическими точками* (границами) $\kappa_{\text{кр.}}$, при этом различают одностороннюю (право- или левостороннюю) и двустороннюю критические области.

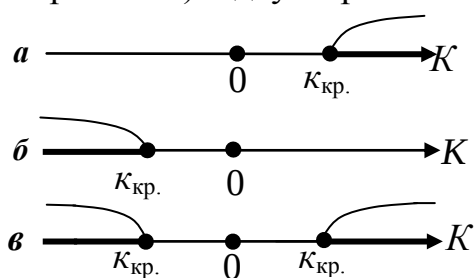


Рис. 13

Правосторонней называют критическую область, определяемую неравенством $K > \kappa_{\text{кр.}}$, где $\kappa_{\text{кр.}} > 0$ (рис. 13, а)

Левосторонней называют критическую область, определяемую неравенством $K < \kappa_{\text{кр.}}$, где $\kappa_{\text{кр.}} < 0$ (рис. 13, б).

Двусторонней называют критическую область, которая определяется неравенствами $K < \kappa_1$, $K > \kappa_2$, где $\kappa_2 > \kappa_1$. В частности, если критические точки симметричны относительно нуля, т. е. распределение критерия симметрично относительно нуля, то двусторонняя критическая область определяется неравенствами (в предположении, что $\kappa_{\text{кр.}} > 0$): $K < -\kappa_{\text{кр.}}$, $K > \kappa_{\text{кр.}}$ или $|K| > \kappa_{\text{кр.}}$ (рис 13, в).

Каким же образом можно отыскать критическую точку $\kappa_{\text{кр.}}$ на оси значений K ? Обычно задают вероятность отклонения гипотезы H_0 , когда она верна. Эта вероятность определяется выбранным уровнем значимости α , уже введенным нами ранее (см. подразд. 2.5), и здесь она называется *уровнем значимости критерия*. Обычно $\alpha = 0,05$, $0,01$ или $0,001$. Если, например, принят уровень значимости равный $0,05$, то это означает, что в 5 случаях из 100 мы рискуем допустить ошибку — отвергнуть правильную гипотезу.

В статистике, принимая решение по результатам проверки гипотезы, можно допустить ошибки различного характера. *Ошибка первого рода* состоит в том, что с вероятностью α отклоняется правильная гипотеза H_0 . *Ошибка второго рода* состоит в том, что будет принята гипотеза H_0 , в то время как она не верна. Вероятность ошибки второго рода обозначают β . Величину $1 - \beta$ называют *мощностью критерия*. Фактически мощность критерия определяется вероятностью не допустить ошибку второго рода ($\beta \rightarrow 0$). Чем ближе мощность критерия к единице, тем более эффективен критерий. Многие статистические критерии получены путем нахождения наиболее мощного критерия при заданных предположениях об основной и альтернативной гипотезах.

Вернемся к вероятности α . Она численно равна площади под кривой распределения критерия K , которая соответствует критической области (например, заштрихованная область на рис. 14).

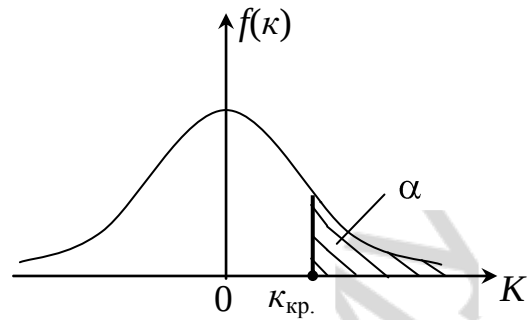


Рис. 14

Для каждого критерия существуют соответствующие таблицы (см. например (4, 5, 9)), по которым при заданном уровне значимости α находят $k_{кр.}$, соответствующие расчетные функции имеются в программах обработки статистических данных, в том числе в табличном процессоре Excel.

Итак, если вычисленное по выборке $K_{набл.}$ попадает в критическую область, нулевая гипотеза H_0 отвергается, если нет, то нет оснований ее отвергнуть.

В современных статистических пакетах обычно сравниваются не только $K_{набл.}$ и $k_{кр.}$, но и заданный уровень значимости α и вероятность того, что, например, для правосторонней критической области $K > K_{набл.}$. Обозначим эту вероятность P , в нашем примере она равна площади под кривой распределения критерия, расположенной справа от $K_{набл.}$ (рис. 15, а, б).

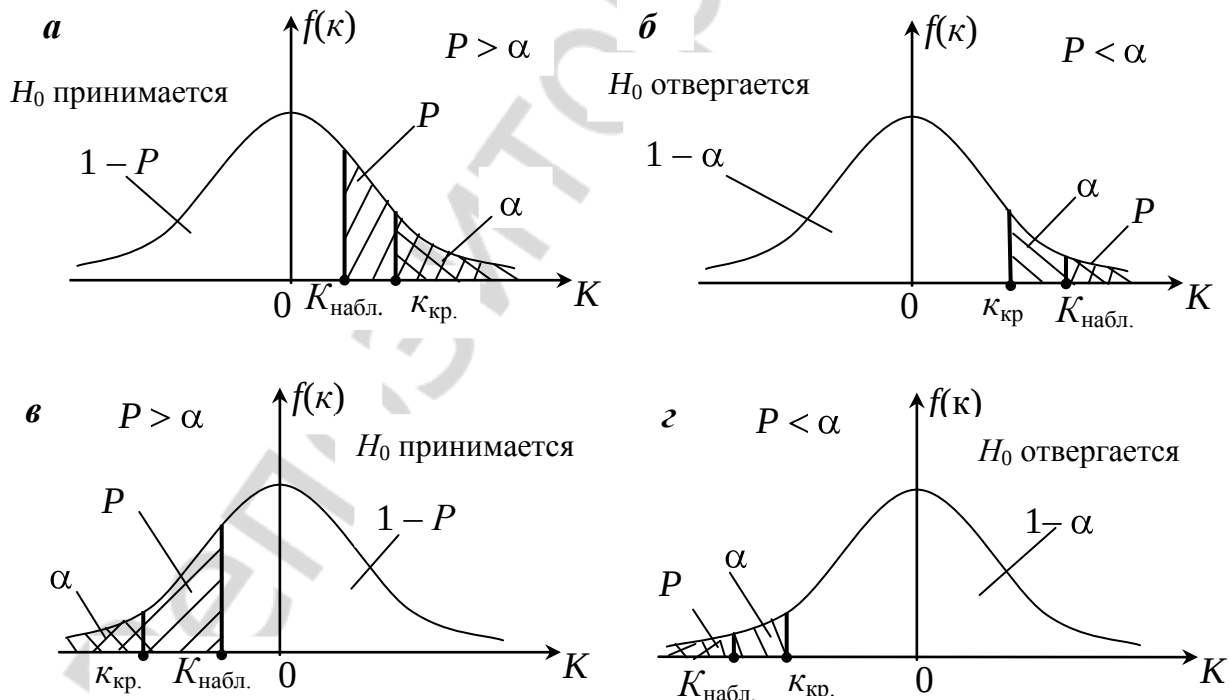


Рис. 15

Если вероятность P оказывается больше заданного уровня значимости α ($P > \alpha$), то гипотеза H_0 принимается (рис 15, а), в противном случае — не принимается (рис 15, б, $P < \alpha$). Рис. 15, в, г иллюстрирует такой подход для левосторонней критической области. Этот подход также верен для двусторонней симметричной критической области (рис. 16).

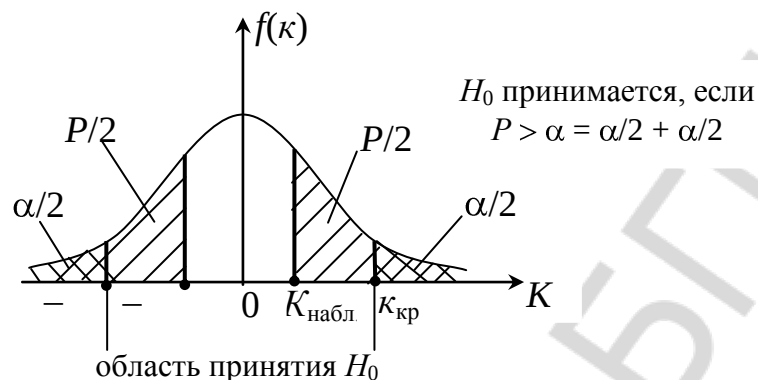


Рис. 16

Отметим, что в данном учебно-методическом пособии, рассматривая примеры использования различных критериев, мы ограничиваемся проверкой гипотез, связанных с параметрами нормального распределения.

2.7. ПРИМЕРЫ РАЗЛИЧНЫХ КРИТЕРИЕВ И ПРАВИЛА РАБОТЫ С НИМИ

1. Работа с одной выборкой.

Проверка гипотезы о значении математического ожидания случайной величины, распределенной по нормальному закону (одновыборочный t-критерий Стьюдента).

Постановка задачи: пусть получена выборка $x_1, x_2, x_3, \dots, x_n$ из нормальной генеральной совокупности случайной величины X . Математическое ожидание (среднее) генеральной совокупности неизвестно, но есть основание предполагать, что $M(X)$ равно некоторому конкретному числу A , например, какому-то стандартному значению.

Требуется проверить гипотезу $H_0: M(X) = A$, исходя из полученной выборки.

Для проверки H_0 необходимо учитывать дисперсию, т. е. степень разброса значений X относительно среднего. В связи с этим рассматриваются 2 случая: дисперсия генеральной совокупности известна (большие выборки) и дисперсия генеральной совокупности неизвестна (малые выборки, $n < 30$). Выберем второй случай, как наиболее практически значимый.

Здесь в качестве критерия (статистики) используют одновыборочный критерий Стьюдента с $(n - 1)$ степенями свободы¹:

$$t = \frac{\bar{x}_в - A}{S} \sqrt{n}, \quad (26)$$

где $\bar{x}_в$ и S — выборочные среднее и стандартное отклонение, n — объем выборки.

Отметим, что при решении соответствующих задач с использованием программы Excel формула (26) вводится с клавиатуры.

Подставляя в (26) заданные A и n и найденные по выборке числовые значения $\bar{x}_в$ и S , вычисляют наблюдаемое (расчетное) значение критерия $t_{набл.}$.

При получении конкретного вывода о принятии либо нет нулевой гипотезы H_0 придерживаются следующих правил, которые определяются видом конкурирующей гипотезы.

1. Если конкурирующая (альтернативная) гипотеза $H_1: M(X) \neq A$ (это значит $M(x) < A$ и $M(x) > A$, т. е. двухсторонняя критическая область), то по заданному уровню значимости α и числу степеней свободы $m = n - 1$ находят критическую точку $t_{двухст. кр.}(\alpha, m)$. Если $|t_{набл.}| < t_{двухст. кр.}$ — нет оснований отвергать нулевую гипотезу, H_0 принимается с заданным уровнем значимости. Если $|t_{набл.}| > t_{двухст. кр.}$ — H_0 отвергают.

При известной вероятности P нулевую гипотезу (H_0) принимают при $P > \alpha$ и не принимают, если $P < \alpha$.

2. Если конкурирующая гипотеза $H_1: M(X) > A$ (правосторонняя критическая область), то по заданным α и m находят $t_{прав. кр.}(\alpha, m)$.

При $t_{набл.} < t_{прав. кр.}$ H_0 принимают, в противном случае — отвергают и принимается конкурирующая гипотеза.

При известной вероятности P нулевую гипотезу (H_0) принимают при $P > \alpha$ и не принимают, если $P < \alpha$.

3. Если конкурирующая гипотеза $H_1: M(X) < A$ (левосторонняя критическая область), то поступают следующим образом. Вначале находят «вспомогательную» критическую точку $t_{правост. кр.}(\alpha, m)$ и полагают границу левосторонней критической области $t_{левост. кр.} = -t_{правост. кр.}$. Если $t_{набл.} > -t_{прав. кр.}$, нет оснований отвергать H_0 , если $t_{набл.} < -t_{правост. кр.}$ H_0 отвергают. Так же как в предыдущих случаях H_0 принимают при $P > \alpha$.

Следует отметить, что данный критерий позволяет провести сравнение выборочной средней с предполагаемой генеральной средней нормальной совокупности. Его часто называют *критерием сравнения выборочной средней с гипотетической генеральной средней*. Такая возмож-

¹ Степени свободы — специальные параметры (характеристики распределения), используемые при работе со статистическими гипотезами.

ность важна при решении многих прикладных задач, в том числе возникающих в медицинской промышленности.

Если H_0 принимается, т. е. $M(X) = A$, то выборочная средняя $\bar{x}_v$ также незначимо отличается от гипотетической генеральной средней A , а если справедлива гипотеза H_1 , то различие этих величин значимо.

Пример. Проектный размер изделий, изготавливаемых станком-автоматом, $A = 35$ мм. Предприятием получен станок, требует ли он корректировки в конкретных условиях работы, если измерения 20 случайно отобранных изделий дали следующие результаты: $\bar{x}_v = 35,07$ мм, $S = 0,16$ мм, $t_{\text{набл.}} = 1,96$?

Проверим при $\alpha = 0,05$ основную гипотезу $H_0: M(X) = 35$ при $H_1: M(X) \neq 35$.

Найденное $t_{\text{двухст. кр.}}(0,05; 19) = 2,09$ и так как $t_{\text{набл.}} < t_{\text{двухст. кр.}}$, то нет оснований при уровне значимости 0,05 отвергать $H_0: M(X) = 35$ мм, так что отличие $\bar{x}_v$ от $A = 35$ мм тоже незначимо. Станок обеспечивает проектный размер изделий и не требует корректировки своей работы.

В заключение отметим, что рассмотренный критерий нечувствителен к умеренным отклонениям от предположения о нормальности распределения.

Важно так же следующее: для проверки гипотезы $H_0: M(X) = A$ против любой из возможных альтернатив H_1 можно использовать доверительный интервал. Мы отвергаем H_0 с уровнем значимости α , если A лежит вне определенного с доверительной вероятностью $\gamma = 1 - \alpha$ доверительного интервала.

Пример. Средний рост младенцев в нормально распределенной популяции новорожденных составляет 51,35 см ($A = 51,35$ см). По данным выборки (здесь мы ее не приводим), предоставленной одним из родильных домов, средний рост новорожденных мальчиков $\bar{x}_v = 51,8$ см, выборочная дисперсия $S^2 = 2,1$ см², объем выборки $n = 25$.

Можно предположить, что средний рост $M(X)$ в популяции новорожденных мальчиков больше чем 51,35 см. Чтобы подтвердить либо опровергнуть это предположение при уровне значимости $\alpha = 0,05$, проверим гипотезу $H_0: M(X) = 51,35$ см против альтернативы $H_1: M(X) > 51,35$ см (правосторонняя критическая область). Вычислим $t_{\text{набл.}}$ по (26):

$$t_{\text{набл.}} = \frac{51,8 - 51,35}{1,45} \cdot 5 = 1,55$$

и найдем $t_{\text{правост. кр.}}; t_{\text{правост. кр.}} = 1,71$.

Так как $t_{\text{набл.}} < t_{\text{правост. кр.}}$ с заданным уровнем значимости $\alpha = 0,05$ принимается H_0 . Это подтверждается также соотношением между P и α : вычисленное $P = 0,067$, значит $P > \alpha$.

Вывод: средний рост новорожденных мальчиков $M(X)$ и $\bar{x}_B$ незначимо отличаются от среднего роста новорожденных младенцев.

Рассчитаем по данным выборки доверительный интервал для $M(X)$, коэффициент Стьюдента $t_{0,95;25} = 2,1$. Расчет дает: полуширина интервала $\delta = 0,61$ см, а $51,19 \text{ см} < M(X) < 52,41 \text{ см}$.

Вывод: этот расчет подтверждает справедливость H_0 : с вероятностью 95 % полученный доверительный интервал содержит и средний рост новорожденных $A = 51,35$ см.

II. Работа с двумя выборками.

Для работы с рассмотренными ниже критериями в программе Excel имеется инструмент «Анализ данных».

1. Проверка гипотезы о равенстве математических ожиданий (средних) двух нормальных генеральных совокупностей при неизвестных, но одинаковых дисперсиях¹ (малые независимые выборки², двухвыборочный t -критерий Стьюдента).

Постановка задачи: получены 2 независимые выборки из нормальных генеральных совокупностей случайных величин X и Y , их объемы n_1 и n_2 . По этим выборкам найдены $\bar{x}_B$, $\bar{y}_B$, S_x^2 и S_y^2 .

Требуется проверить гипотезу $H_0: M(X) = M(Y)$.

Здесь в качестве критерия (статистики) используется двухвыборочный критерий Стьюдента:

$$t = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{(n_1 - 1)S_x^2 + (n_2 - 1)S_y^2} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}}. \quad (28)$$

При $n_1 = n_2 = n$

$$t = \frac{\bar{x}_B - \bar{y}_B}{\sqrt{S_x^2 + S_y^2}} \sqrt{n} \quad (29)$$

Дальнейшие действия стандартны. Необходимо вычислить $t_{\text{набл.}}$ с помощью формулы (28) или (29) и конкретных характеристик выборок. Если полученное значение $t_{\text{набл.}}$ принадлежит критической области, то H_0 отвергается и принимается конкурирующая гипотеза $H_1: M(X) \neq M(Y)$, а следовательно, $\bar{x}_B$ и $\bar{y}_B$ различаются значимо, т. е. их различие вызвано принципиальными причинами. Если $t_{\text{набл.}}$ оказывается в области принятия нулевой гипотезы, то $M(X) = M(Y)$ и различие выборочных средних незначимо и обусловлено случайными факторами.

¹ Равенство дисперсий можно проверить, используя критерий Фишера–Снедекора, который рассматривается ниже.

² Независимые выборки — выборки, полученные для разных объектов, связанных определенным исследованием.

Правила, позволяющие сделать вывод о справедливости нулевой гипотезы, зависят от конкурирующей гипотезы H_1 , от уровня значимости α и числа степеней свободы $m = n_1 + n_2 - 2$ или $m = 2(n - 1)$ (при $n_1 = n_2 = n$). Они не отличаются от рассмотренных в первом случае работы с одной выборкой.

Если рассчитывается вероятность P , то H_0 принимается при $P > \alpha$.

Пример. Сравнительное исследование концентрации свинца в крови (в мг/100 г) группы рабочих аккумуляторного завода X (подвергавшихся профессиональному воздействию) и группы рабочих текстильной фабрики Y (не подвергавшихся профессиональному воздействию), привело к следующим результатам:

$$\bar{x}_в = 0,08157 \text{ мг/100 г}, S_x = 0,0067 \text{ мг/100 г}, S_x^2 = 4,489 \cdot 10^{-5}, n = 7$$

$$\bar{y}_в = 0,03943 \text{ мг/100 г}, S_y = 0,00355 \text{ мг/100 г}, S_y^2 = 1,26 \cdot 10^{-5}, n = 7.$$

Число степеней свободы $m = 12$.

Предполагается, что $D(X) = D(Y)$ и исследуемый показатель в генеральной совокупности распределен по нормальному закону.

При $\alpha = 0,05$ проверяется $H_0: M(X) = M(Y)$ против альтернативы $H_1: M(X) \neq M(Y)$. В соответствии с вышеприведенными числовыми данными $t_{\text{набл.}} = 19,6$, $t_{\text{двухст. кр.}} = 2,18$.

Так как $t_{\text{набл.}} > t_{\text{двухст. кр.}}$, нулевая гипотеза отвергается с заданным уровнем значимости.

То же подтверждает расчет P , $P < 0,05$.

Вывод: условия работы значимо влияют на содержание свинца в крови рабочих.

2. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей (F-критерий Фишера–Снедекора).

Постановка задачи: пусть генеральные совокупности величин X и Y распределены по нормальному закону. По независимым выборкам объемами n_1 и n_2 , извлеченным из этих совокупностей, найдены выборочные дисперсии S_x^2 и S_y^2 . По этим дисперсиям при заданном уровне значимости α требуется проверить нулевую гипотезу, что генеральные дисперсии рассматриваемых совокупностей равны между собой:

$$H_0: D(X) = D(Y).$$

Если окажется, что гипотеза H_0 справедлива, т. е. генеральные дисперсии одинаковы, то различие выборочных дисперсий незначимо и объясняется случайными причинами. Если H_0 будет отвергнута, т. е. если генеральные дисперсии неодинаковы, то различие выборочных дисперсий значимо. Оно не может быть объяснено случайными причинами, а является следствием различия самих генеральных дисперсий.

В качестве критерия (статистики) проверки нулевой гипотезы о равенстве генеральных дисперсий принимают величину:

$$F = \frac{S_6^2}{S_M^2} \text{ (критерий } F \text{ Фишера–Снедекора),} \quad (30)$$

где S_M и S_6 — соответственно меньшее и большее значение выборочных дисперсий.

При использовании критерия (30) критическая область, как всегда, определяется видом конкурирующей гипотезы. Рассмотрим два случая.

1. Нулевая гипотеза $H_0: D(X) = D(Y)$.

Конкурирующая гипотеза $H_1: D(X) \neq D(Y)$. Здесь критическая область *двусторонняя*.

2. Если есть основание предполагать, что одна из дисперсий обязательно не меньше другой, например, $D(X) \geq D(Y)$, тогда нулевая гипотеза $H_0: D(X) = D(Y)$, а конкурирующая гипотеза $H_1: D(X) > D(Y)$. В данном случае критическая область *правосторонняя*. Если $H_1: D(X) < D(Y)$, критическая область *левосторонняя*.

При решении конкретных задач придерживаются следующих правил:

1. Чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(X) = D(Y)$ при конкурирующей гипотезе $H_1: D(X) > D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей

выборочной дисперсии к меньшей): $F_{\text{набл.}} = \frac{S_6^2}{S_M^2} \geq 1$

Далее по заданному α и числам степеней свободы $m_1 = n_1 - 1$ и $m_2 = n_2 - 1$ (n_1 — объем выборки, для которой получена большая выборочная дисперсия) по соответствующим таблицам находят критическую точку $F_{\text{кр.}}(\alpha, m_1, m_2)$. Если $F_{\text{набл.}} < F_{\text{кр.}}$ — нет оснований отвергать H_0 , и она принимается. Если же $F_{\text{набл.}} > F_{\text{кр.}}$, то H_0 отвергают и полагают, что $D(X) > D(Y)$.

При работе со статистическими пакетами гипотезу H_0 принимают, если $P > \alpha$, в противном случае — нет.

2. При конкурирующей гипотезе $H_1: D(X) \neq D(Y)$ критическую точку $F_{\text{кр.}}(\alpha/2, m_1, m_2)$ ищут по уровню значимости $\alpha/2$. Если $F_{\text{набл.}} < F_{\text{кр.}}$ — нет оснований отвергать H_0 . Если $F_{\text{набл.}} > F_{\text{кр.}}$ — H_0 отвергают.

При известной вероятности P гипотезу H_0 принимают, если $P > \alpha = \alpha/2 + \alpha/2$.

Пример. Условие задачи и соответствующие данные приведены в примере, который иллюстрирует работу двухвыборочного критерия Стьюдента. Воспользуемся ими в данном случае.

Сформулируем вопрос: при уровне значимости $\alpha = 0,05$ проверить $H_0: D(X) = D(Y)$, при $H_1: D(X) \neq D(Y)$.

По данным задачи, $F_{\text{набл.}} = \frac{S_x^2}{S_y^2} = 3,56$. Так как критическая область

двусторонняя, то при отыскании критической точки следует брать уровень значимости в два раза меньший заданного, то есть $\alpha/2 = 0,025$, тогда $F_{\text{кр.}} = (0,025; 6; 6) = 5,82$.

Так как $F_{\text{набл.}} < F_{\text{кр.}}$, H_0 принимается — $D(X) = D(Y)$.

3. Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки¹).

Постановка задачи: пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии неизвестны. Требуется, используя зависимые выборки, при уровне значимости α , проверить основную гипотезу $H_0: M(X) = M(Y)$ при альтернативе $H_1: M(X) \neq M(Y)$ или $H_1: M(X) > M(Y)$, или $H_1: M(X) < M(Y)$.

В этом случае используется парный двухвыборочный t -критерий Стьюдента с $m = n - 1$ степенями свободы, который имеет вид:

$$t = \frac{\bar{d}}{S_d} \cdot \sqrt{n}, \quad (31)$$

где $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$, $d_i = x_i - y_i$ ($i = 1 \dots n$), $S_d = \sqrt{\sum_{i=1}^n \frac{(d_i - \bar{d})^2}{n-1}}$, n — объем выборки.

Далее решение задачи находится тривиально: сравниваем $t_{\text{набл.}}$, полученное по данным выборки из (31), и $t_{\text{кр.}}$ при разных вариантах критических областей.

Например, если $H_1: M(X) \neq M(Y)$ и $|t_{\text{набл.}}| < t_{\text{двухст. кр.}}$, H_0 принимается, в противном случае принимается H_1 .

При возможности вычисления P , H_0 принимается при $P > \alpha$.

Пример. Рассмотрим популяцию, состоящую из критически больных пациентов с циркуляторным шоком. Была получена выборка из 108 пациентов и у каждого из них измерялось X — венозное рН и Y — артериальное рН.

Из клинического опыта известно, что для здоровых людей среднее венозное рН меньше, чем артериальное. Проверим, выполняется ли это соотношение для популяции больных с указанной выше патологией, используя парный t -критерий Стьюдента, примем $H_0: M(X) = M(Y)$, а $H_1: M(X) < M(Y)$.

¹ Зависимые выборки — выборки, полученные для одних и тех же объектов, связанных определенным исследованием.

По литературным данным [1], $\bar{d} = -0,04$, $S_d = 0,1533$, $\sqrt{n} = \sqrt{108} = 10,39$. Используя (31), получим, что $t_{\text{набл.}} = -2,71$, $t_{\text{левост. кр.}} = -1,66$ на уровне значимости $\alpha = 0,05$. Так как $t_{\text{набл.}} < t_{\text{левост. кр.}}$, то H_0 отвергается и принимается H_1 .

Вывод: в популяции критически больных пациентов среднее венозное рН и среднее артериальное рН значительно отличаются друг от друга, причем $(\text{pH})_{\text{вен.}} < (\text{pH})_{\text{арт.}}$. Это неравенство подтверждается медицинскими фактами.

2.8. ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

Одной из задач анализа данных является установление зависимости (связи) между признаками — случайными величинами (частота пульса, артериальное давление, показатель анализа крови и т. д.). Эта задача решается методами корреляционного¹ анализа.

Пусть X и Y — случайные величины. Зависимость их друг от друга (если она существует) называется корреляционной зависимостью. Она может быть установлена качественно — по виду диаграммы рассеяния (корреляционного поля), и количественно — путем вычисления коэффициента корреляции. При установлении корреляционной зависимости экспериментально для каждого обследованного объекта получают соответствующие пары значений величин X и Y (например, роста и массы тела людей определенного пола и возраста, числа эритроцитов и содержания гемоглобина в анализе крови).

Пусть объем выборки — n . Каждой паре значений (x_i, y_i) на плоскости xOy соответствует одна точка. Всего будет n точек.

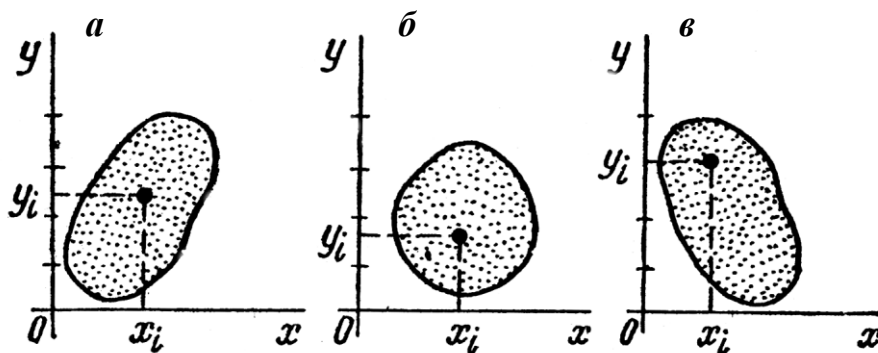


Рис. 17

Область на графике $y(x)$, занятая этими точками, образует диаграмму рассеяния (корреляционное поле). Разные виды таких диаграмм (полей) показаны на рис. 17. Если форма корреляционного поля близка к кругу

¹ Корреляция (от англ. correlation) — согласование, связь, взаимозависимость.

(рис. 17, б), то связи между признаками X и Y нет. Если же корреляционное поле вытянуто (рис. 17, а, 17, в), то корреляционная связь между признаками X и Y есть, и она тем сильнее, чем более вытянуто корреляционное поле.

По экспериментальным данным, для каждого значения признака X можно найти $\bar{Y}$. Зависимость $\bar{Y}_x = f(x)$ называется *эмпирическим уравнением регрессии Y на X* . Аналогично можно получить зависимость $\bar{X}_y = \varphi(y)$ — *уравнение регрессии X на Y* . Графики этих функций называются *линиями регрессии*. Экспериментальные данные более или менее тесно группируются вдоль этих линий. Если они представляют собой прямые, то корреляционная связь между признаками X и Y называется *линейной* и оценивается с помощью *выборочного коэффициента корреляции r* .

Значения r по модулю не превышают 1, но могут быть как положительными, так и отрицательными:

$$-1 \leq r \leq 1 \text{ или } |r| \leq 1.$$

При $r = 0$ линейная связь между X и Y отсутствует; при значениях $|r|$ до 0,3 — связь слабая; от 0,3 до 0,7 — умеренная; от 0,7 до 1 — сильная; если $|r| \approx 1$ — связь полная или, иначе, функциональная — в этом случае существует функция $Y = f(X)$, связывающая значения Y и X .

Вернемся к линиям регрессии. Рассмотрим, например, зависимость $Y = f(X)$. Если ее график — прямая, то ее уравнение ($\bar{Y}_x = f(x)$) можно записать в виде:

$$\bar{Y}_x = kx + b, \quad (31)$$

где постоянные k и b определяются по выборке.

По линиям регрессии можно оценить тенденцию (тренд) изменения одной величины при изменении значений другой. Коэффициент k в (31) называется коэффициентом регрессии. Если $k > 0$, то при увеличении (уменьшении) значений одной величины, например X , увеличиваются (уменьшаются) значения другой Y . При $k < 0$, с увеличением (уменьшением) значений X , значения Y уменьшаются (увеличиваются).

Коэффициент корреляции r имеет тот же знак, что и k . При $r > 0$ связь между признаками X и Y называется *прямой*, при $r < 0$ — *обратной*.

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте зависимости между признаками, полученное по данным выборки, можно распространить и на генеральную совокупность. Например, для оценки коэффициента корреляции r_T нормально распределенной генеральной совокупности (при $n > 50$) можно воспользоваться формулой

$$r - 3\frac{1-r^2}{\sqrt{n}} < r_T < r + 3\frac{1-r^2}{\sqrt{n}}.$$

Глава 3

АНАЛИЗ ДАННЫХ С ПОМОЩЬЮ ТАБЛИЧНОГО ПРОЦЕССОРА EXCEL

3.1. ЭТАПЫ ОБРАБОТКИ И АНАЛИЗА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

В соответствии с приведенным в гл. 2 материалом определим этапы обработки и анализа экспериментальных данных. Здесь необходимо:

1. Получить выборку объема n .
 2. Составить ряд распределения (вариационный или интервальный).
 3. Проиллюстрировать ряд распределения графически (построить полигон частот и/или гистограмму).
 4. Определить числовые характеристики выборки (среднее арифметическое, дисперсию, стандартное отклонение, стандартную ошибку среднего, моду, медиану, коэффициенты асимметрии, эксцесса и вариации).
 5. По опытным данным сделать вывод о законе распределения исследуемого показателя в генеральной совокупности.
 6. Найти доверительные интервалы для параметров распределения.
 7. При необходимости, выбрав соответствующий критерий, провести проверку гипотезы, связанной с параметрами нормального распределения.
- Покажем реализацию этих этапов, используя конкретные примеры.

3.2. ОПИСАТЕЛЬНАЯ СТАТИСТИКА. ОПРЕДЕЛЕНИЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ВЫБОРКИ С ПОМОЩЬЮ ФОРМУЛ И МАСТЕРА ФУНКЦИЙ

Основные вопросы:

1. Ввод данных в таблицу.
2. Расчет выборочных характеристик случайной величины с помощью Мастера функций Excel.
3. Расчет выборочных характеристик с помощью формул.

Рассмотрим, взяв конкретный пример, расчет числовых характеристик случайной величины, который в статистическом анализе данных традиционно относят к разделу «Описательная статистика», с помощью мастера функций Excel.

Задача. Анализируемый показатель — концентрация сывороточного альбумина (г/л), содержащегося в сыворотке крови женщин. В результате исследований получена следующая выборка значений этого показателя:

42	41	42	44	44	36	38	41	42	44	42	39	49	40	45	32	34	43	37
39	41	39	48	42	43	33	43	35	32	34	39	35	43	44	47	40	39	42
41	46	37	49	41	39	43	42	47	48	51	52							

Определите следующие статистические характеристики: выборочные среднее, медиану, моду, дисперсию, стандартное отклонение, максимальное и минимальное значение концентрации, объем выборки, вариационный размах выборки, стандартную ошибку среднего, коэффициент асимметрии и эксцесс.

3.2.1. СОЗДАНИЕ КНИГИ. ВВОД ДАННЫХ В ТАБЛИЦУ

Для ввода данных в таблицу выполните следующие действия:

1. Запустите пакет Excel, сохраните созданный файл под именем **Статистика**.

2. Переименуйте **Лист1** в **Описательная статистика**.

При расчете статистических характеристик с помощью **Мастера функций** исходные числовые данные можно представлять в виде прямоугольной области, вертикального столбца или горизонтальной строки. Если используется выборка значений только одной случайной величины, удобнее прямоугольная область.

3. Введите с помощью клавиатуры данные задачи в таблицу в соответствии с рис. 18. Для этого:

– введите в ячейку **A1** название исследуемого параметра «Содержание сывороточного альбумина»;

– заполните числовыми данными блок ячеек таблицы **A2:E11**.

4. Столбец **G** отведите названиям статистических характеристик исследуемой величины. Для этого:

– в ячейку **G1** введите заголовок «Описательная статистика»;

– в ячейки **G2:G14** введите названия вычисляемых характеристик (рис. 18).

5. Увеличьте ширину столбца **G** следующим образом: выделите щелчком мыши столбец G, затем дважды щелкните правую границу заголовка столбца.

	A	B	C	D	E	F	G
	Содержание сывороточного альбумина						Описательная статистика
1							
2	42	41	42	44	44		Среднее
3	36	38	41	42	44		Мода
4	42	39	49	40	45		Медиана
5	32	34	43	37	39		Минимум
6	41	39	48	42	43		Максимум
7	33	43	35	32	34		Объем выборки
8	39	35	43	44	47		Асимметрия
9	40	39	42	41	46		Эксцесс
10	37	49	41	39	43		Дисперсия
11	42	47	48	51	52		Стандартное отклонение
12							Вариационный размах
13							Стандартная ошибка

Рис. 18

3.2.2. РАСЧЕТ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ С ПОМОЩЬЮ МАСТЕРА ФУНКЦИЙ EXCEL

Вычисление числовых значений следующих характеристик: выборочного среднего, медианы, моды, дисперсии выборочного среднего, стандартного отклонения, максимального и минимального значений концентрации, объема выборки, коэффициента вариации, коэффициента асимметрии и эксцесса — производится с помощью **Мастера функций**.

Соответствие между названиями вычисляемых характеристик выборки и их обозначением в **Мастере функций** приведено в табл. 3.

Таблица 3

Характеристики положения	Среднее выборочное	СРЗНАЧ
	Мода	МОДА.ОДН
	Медиана	МЕДИАНА
	Минимум	МИН
	Максимум	МАКС
Характеристики разброса	Дисперсия	ДИСП.В
	Стандартное отклонение	СТАНДОТКЛОН.В
	Размах выборки	нет встроенной функции
Характеристики формы	Коэффициент асимметрии	СКОС
	Эксцесс	ЭКСЦЕСС
	Объем выборки	СЧЕТ

Расчет значения любой из характеристик требует определенной последовательности действий. Разберем ее на следующем примере:

1. Для вычисления значения выборочного среднего:

– выделите щелчком мыши ячейку **H2**;

– нажмите кнопку 

Вставка функции;

– в появившемся окне **Мастер функций** (рис. 19) в поле *Категория* щелчком мыши выберите **Статистические**;

– в поле *Выберите функцию*, листая список названий функций, найдите и выделите щелчком функцию **СРЗНАЧ**; подтвердите выбор, нажав **ОК**;

– в появившемся окне **Аргументы функции** (рис. 20) в поле *Число1* вручную введите адрес диапазона ячеек с данными **A2:E11** или выделите этот диапазон в таблице мышью; подтвердите, нажав **ОК**;

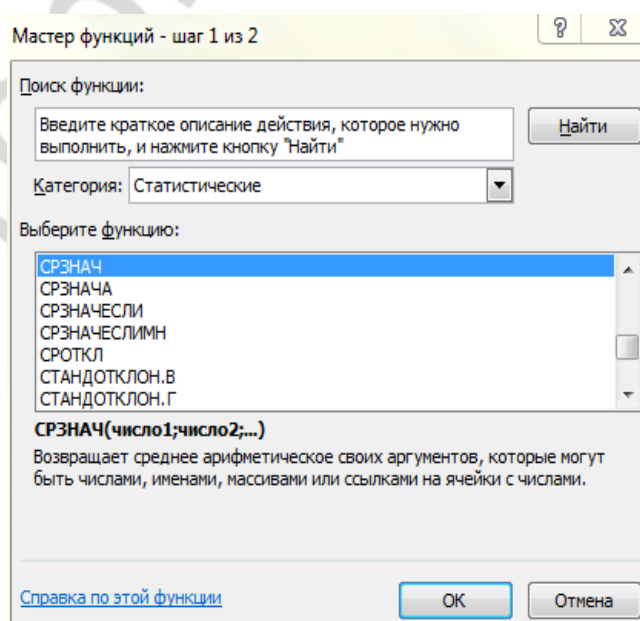


Рис. 19

После этого в ячейке **Н2** появится результат расчета среднего значения.

	A	B	C	D	E	F	G	H
	Содержание сывороточного							
1	альбумина					Описательная статистика		
2	42	41	42	44	44	Среднее	=СРЗНАЧ(A2:E11)	
3	36	38	41	42	44	Мода		
4	42							
5	32							
6	41							
7	33							
8	39							
9	40							
10	37							
11	42							
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								

Аргументы функции

СРЗНАЧ

Число1

A2:E11

Число2

= {42;41;42;44;44;36;38;41;42;44;2;...}

= число

= 41,38

Возвращает среднее арифметическое своих аргументов, которые могут быть числами, именами, массивами или ссылками на ячейки с числами.

Число1: число1;число2;... от 1 до 255 числовых аргументов, для которых вычисляется среднее.

Значение: 41,38

[Справка по этой функции](#)

OK

Отмена

Рис. 20

2. Аналогично выбирая соответствующие функции, вычислите в ячейке **Н3** значение моды, в ячейке **Н4** значение медианы, в ячейках **Н5** и **Н6** значения **МИН** и **МАКС** и т. д.

Таким же образом рассчитайте значения остальных характеристик: асимметрии, эксцесса, дисперсии и стандартного отклонения, используя для этого функции из категории **Статистические** соответственно табл. 3.

3.2.3. РАСЧЕТ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ С ПОМОЩЬЮ ФОРМУЛ

Чтобы рассчитать вариационный размах, стандартную ошибку выборочного среднего, необходимо вводить формулы вручную, т. к. для этих характеристик встроенные функции в Мастере не предусмотрены.

Для расчета вариационного размаха нужно вычислить $(x_{\text{макс}} - x_{\text{мин}})$ разность между максимальным и минимальным значением выборки, для расчета стандартной ошибки $(S/\sqrt{n})$ — разделить стандартное отклонение на корень из объема выборки.

Чтобы *рассчитать вариационный размах* введите в ячейку **Н12** следующую формулу: **=Н6-Н5**.

Для *вычисления стандартной ошибки выборочного среднего* установите курсор в ячейку **Н14** и введите в нее формулу **=Н11/КОРЕНЬ(Н7)**:

– для вызова функции **КОРЕНЬ** нажмите кнопку  **Вставка функции**;

– в появившемся окне **Мастер функций** в поле *Категория* щелчком мыши выберите **Математические**;

– в поле *Выберите функцию*, найдите и выделите щелчком функцию **КОРЕНЬ**, подтвердите выбор, нажав **ОК**;

– в появившемся окне **Аргументы функции** в поле *Число* укажите адрес ячейки **Н7**, подтвердите, нажав **ОК**.

Окончательный результат ваших действий должен соответствовать рис. 21.

	Г	Н		Г	Н
1	Описательная статистика		1	Описательная статистика	
2	Среднее	=СРЗНАЧ(A2:E11)	2	Среднее	41,38
3	Мода	=МОДА.ОДН(A2:E11)	3	Мода	42
4	Медиана	=МЕДИАНА(A2:E11)	4	Медиана	42
5	Минимум	=МИН(A2:E11)	5	Минимум	32
6	Максимум	=МАКС(A2:E11)	6	Максимум	52
7	Объем выборки	=СЧЁТ(A2:E11)	7	Объем выборки	50
8	Асимметрия	=СКОС(A2:E11)	8	Асимметрия	0,03
9	Экссесс	=ЭКССЕСС(A2:E11)	9	Экссесс	-0,18
10	Дисперсия	=ДИСП.В(A2:E11)	10	Дисперсия	22,53
11	Стандартное отклонение	=СТАНДОТКЛОН.В(A2:E11)	11	Стандартное отклонение	4,75
12	Вариационный размах	=Н6-Н5	12	Вариационный размах	20
13	Стандартная ошибка	=Н11/КОРЕНЬ(Н7)	13	Стандартная ошибка	0,67

Рис. 21: *а* — в режиме отображения формул; *б* — в режиме отображения результатов расчетов

Сохраните результаты работы в том же файле **Статистика.xlsx** в рабочей папке.

3.2.4. АНАЛИЗ ПОЛУЧЕННЫХ ЧИСЛОВЫХ ХАРАКТЕРИСТИК

На основании проведенных вычислений можно сделать следующие выводы:

1. В данной выборке большая часть значений исследуемого показателя (концентрации сывороточного альбумина) группируется в диапазоне $(41,38 \pm 4,75)$ г/л; центр рассеяния — значение 41,38; наибольшую частоту имеет значение 42.

2. Можно предположить, что в генеральной совокупности исследуемый показатель (концентрация сывороточного альбумина) распределен по нормальному закону, т. к. выборочные значения среднего, моды и медианы незначительно отличаются друг от друга; минимальное и максимальное значения вариант равноудалены от среднего; выборочные Ex_v и As_v близки к нулю.

3.2.5. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. В таблице приведены данные о содержании общего холестерина (ммоль/л) в крови группы пациентов:

7	8	5,4	6,7	7,6	6,6	8	8,4
---	---	-----	-----	-----	-----	---	-----

Определите следующие статистические характеристики для этого показателя: выборочные среднее, медиану, моду, дисперсию, стандартное отклонение, максимальное и минимальное значение, объем выборки, коэффициент асимметрии и эксцесс. Проанализируйте полученные результаты.

Задача № 2. Анализируемый показатель — значение гематокрита (Hct) у больных в критическом состоянии. При поступлении в стационар получены приведенные ниже данные (%).

34	41	46	28	39	30	25	41	42	31	25	20	26	28	27	37	44	41	32	30
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Определите следующие статистические характеристики для этого показателя: выборочные среднее, медиану, моду, дисперсию, стандартное отклонение, объем выборки, коэффициент асимметрии и эксцесс. Сделайте анализ результатов. Можно ли утверждать по этим выборочным данным, что данный показатель генеральной совокупности распределен по нормальному закону?

3.3. ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ СТАТИСТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ ВЫБОРКИ

Основные вопросы:

1. Построение гистограммы частот вручную с помощью Мастера функций.
2. Построение гистограммы относительных частот.
3. Форматирование и редактирование построенной гистограммы.
4. Построение полигона частот.

Гистограмму можно строить пошагово, используя Мастер функций, или одномоментно, используя Пакет анализа. Для построения гистограммы в Мастере функций необходимо выполнить предварительные расчеты: определить число интервалов для данного объема выборки, ширину одного интервала и создать таблицу частот попадания значений случайной величины в заданный интервал. Число интервалов можно определить с помощью формул: $n \approx 5 \cdot \lg N$ или $n \approx \sqrt{N}$, где N — объем выборки, результат округляется до целого значения; ширина интервалов обычно одинакова и равна $h = (\text{МАКС} - \text{МИН})/n$. В задачах данного раздела n будет задано. Напомним, что гистограмму можно строить в разных координатах, откладывая по оси Y частоты для соответствующих интервалов, относительные частоты или плотности этих величин. На оси X всегда указы-

ваются правые границы интервалов (в Пакете анализа они называются карманы).

3.3.1. ПОСТРОЕНИЕ ГИСТОГРАММЫ ЧАСТОТ ВРУЧНУЮ

Задача. Используя данные задачи, приведенной в 3.2, постройте гистограмму частот концентрации альбумина в крови женщин, число интервалов $n \approx 5 \lg 50 \approx 8$.

1. Откройте созданный ранее файл **Статистика.xlsx** с таблицей «Содержание сывороточного альбумина».

2. Введите в ячейки **B15** и **B16** текст в соответствии с рис. 22, в ячейку **D15** введите заданное по условию число интервалов — 8.

	A	B	C	D
14				
15		Количество интервалов		8
16		Ширина интервала		

Рис. 22

3. В ячейку **D16** вставьте формулу для расчета ширины интервала $=(H6 - H5)/D15$, как показано на рис. 23.

	A	B	C	D
14				
15		Количество интервалов	8	
16		Ширина интервала		$=(H6-H5)/D15$

Рис. 23

Результат ваших действий показан на рис. 24.

	A	B	C	D
14				
15		Количество интервалов		8
16		Ширина интервала		2,5

Рис. 24

4. Составьте таблицу частот попадания значений исследуемой величины в заданные интервалы. Для этого введите заголовки столбцов и формулы для расчета границ интервалов и частот как показано на рис. 25. Для вычисления правой границы первого интервала используется уравнение $= \text{МИН} + x$; для остальных интервалов граница определяется путем прибавления величины ширины интервала x к правой границе предыдущего интервала):

	№	Интервалы	Частоты
18			
19	1	$=H5+D16$	
20	2	$=B19+\$D\16	
21	3	$=B20+\$D\16	
22	4	$=B21+\$D\16	
23	5	$=B22+\$D\16	
24	6	$=B23+\$D\16	
25	7	$=B24+\$D\16	
26	8	$=B25+\$D\16	

Рис. 25

– в ячейку **B18** введите заголовок **Интервалы**, в ячейку **C18** — заголовок **Частоты**;

– в блок ячеек **A19:A26** введите номера интервалов с 1 по 8.

5. Вставьте расчетные формулы для вычисления правых границ интервалов в соответствии с рис. 25, для чего:

– в ячейку **B19** введите формулу $=H5+D16$ (минимальное значение + ширина интервала), нажмите **Enter**;

– в ячейку **B20** введите следующую формулу: **B19+D16** (граница первого интервала + ширина интервала), нажмите клавишу **F4**, чтобы сделать ссылку **D16** абсолютной, нажмите **Enter**.

6. Скопируйте формулу из ячейки **B20** при помощи маркера автозаполнения вниз в диапазон **B21:B26**.

7. Вставьте в блок ячеек **C19:C26** функцию для расчета частот, указывая массив данных и массив интервалов¹. Для этого:

– в ячейку **C19** вставьте функцию **ЧАСТОТА** из категории **Статистические**,

– в открывшемся окне **Аргументы функции** (рис. 26) в поле **Массив_данных**, необходимо указать диапазон ячеек исходных данных **A2:E11**;

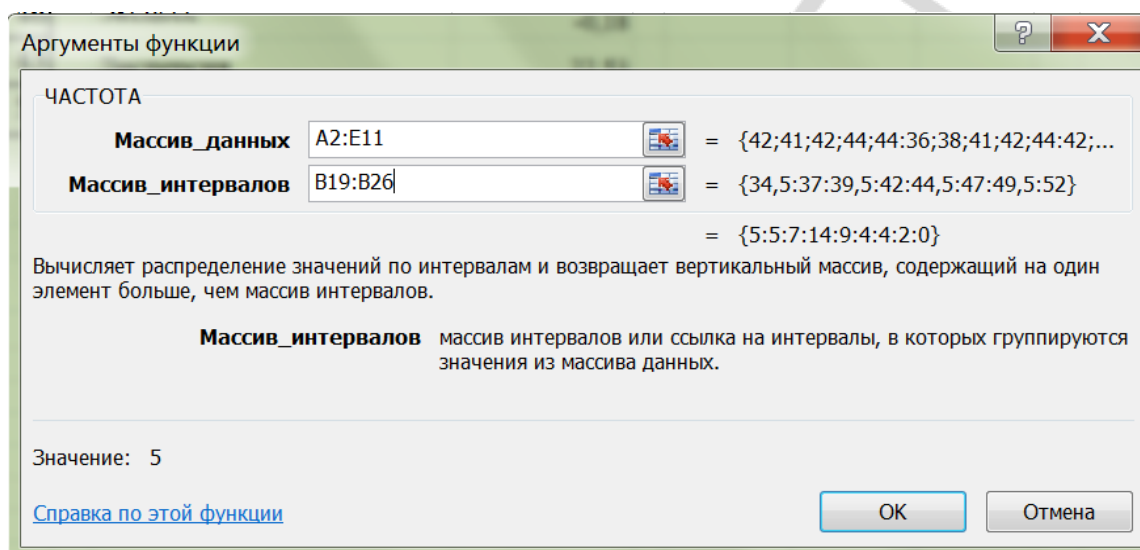


Рис. 26

	A	B	C	D
14				
15		Количество интервалов		8
16		Ширина интервала		2,5
17				
18	№	Интервалы	Частоты	
19	1	34,5	5	
20	2	37	5	
21	3	39,5	7	
22	4	42	14	
23	5	44,5	9	
24	6	47	4	
25	7	49,5	4	
26	8	52	2	

Рис. 27

– в поле **Массив_интервалов** укажите диапазон ячеек, содержащий границы интервалов **B19:B26**, подтвердите, нажав **OK**;

– далее выделите диапазон **C19:C26**, начиная с ячейки, содержащей формулу, нажмите клавишу **F2**, а затем нажмите клавиши **Ctrl + Shift + Enter** для фиксации функции массива.

Результат ваших действий отображен на рис. 27.

¹ Массив данных — набор данных, для которых вычисляются частоты, массив интервалов — набор интервалов, по которым распределяются данные.

8. Постройте гистограмму частот для исследуемой величины, для чего:
- выделите диапазон ячеек **C19:C26**, содержащий значения частот, как показано на рис. 28;
 - на вкладке **Вставка**, в группе **Диаграммы** выберите диаграмму — **Гистограмма**;
 - затем выберите тип диаграммы → **Гистограмма с группировкой**.

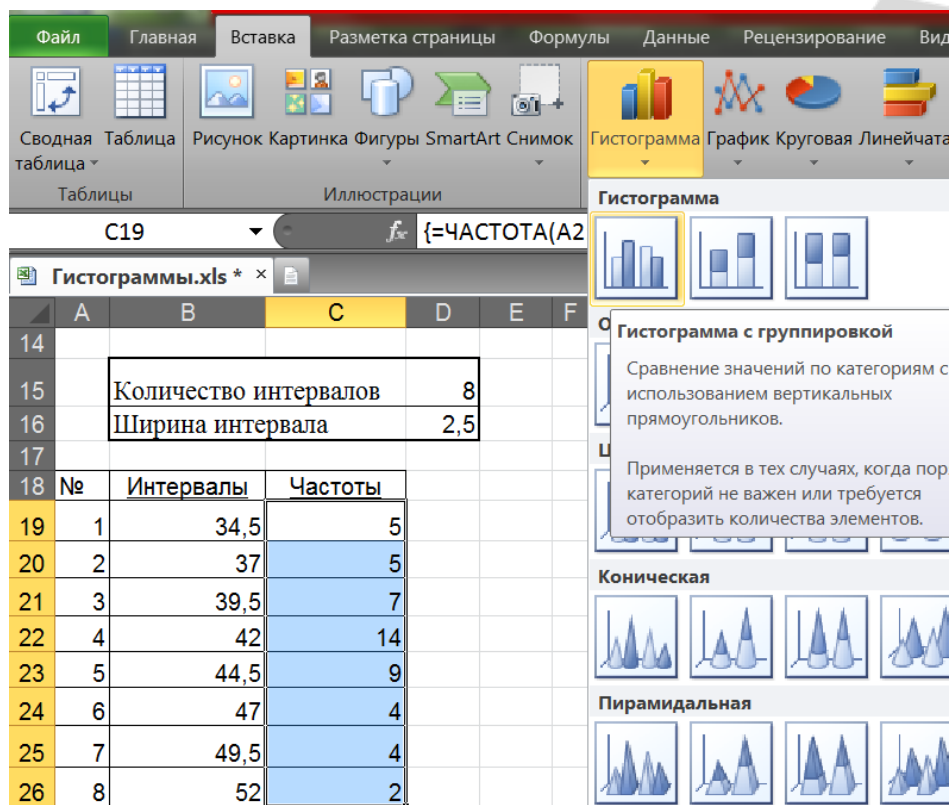


Рис. 28

Результат ваших действий отображен на рис. 29.

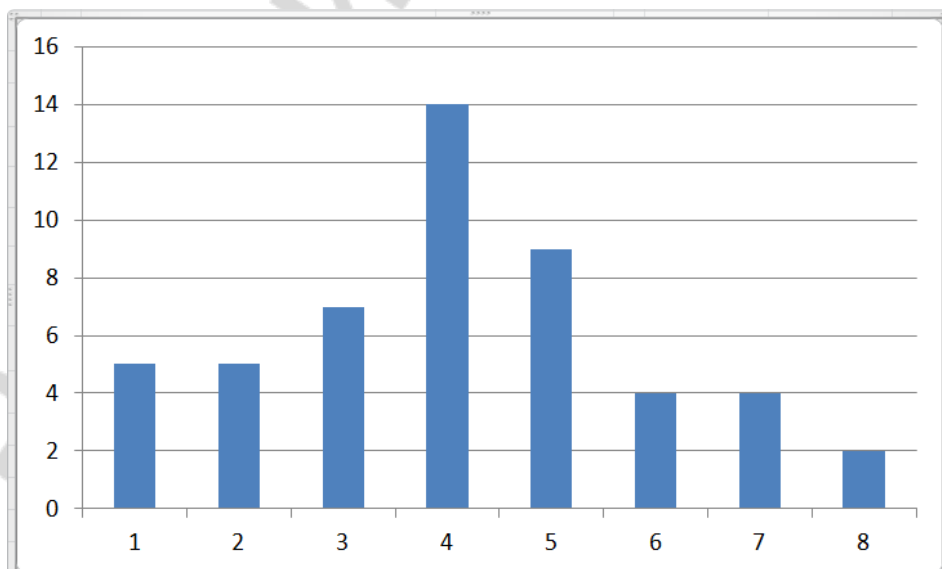


Рис. 29

9. Используя вкладку **Работа с диаграммами/Макет**, дополните построенную гистограмму следующими элементами: **Название диаграммы** и **Названия осей** соответственно рис. 30.



Рис. 30

10. По умолчанию интервалы на гистограмме нумеруются 1, 2, 3 и т. д. Нам же необходимо на горизонтальной оси указать вычисленные ранее правые границы интервалов, как показано на рис. 32. Для этого:

- на вкладке **Работа с диаграммами/Конструктор**, нажмите кнопку **Выбрать данные**;
- в открывшемся окне **Выбор источника данных** (рис. 31) в разделе **Подписи горизонтальной оси (категории)** нажмите кнопку **Изменить**.

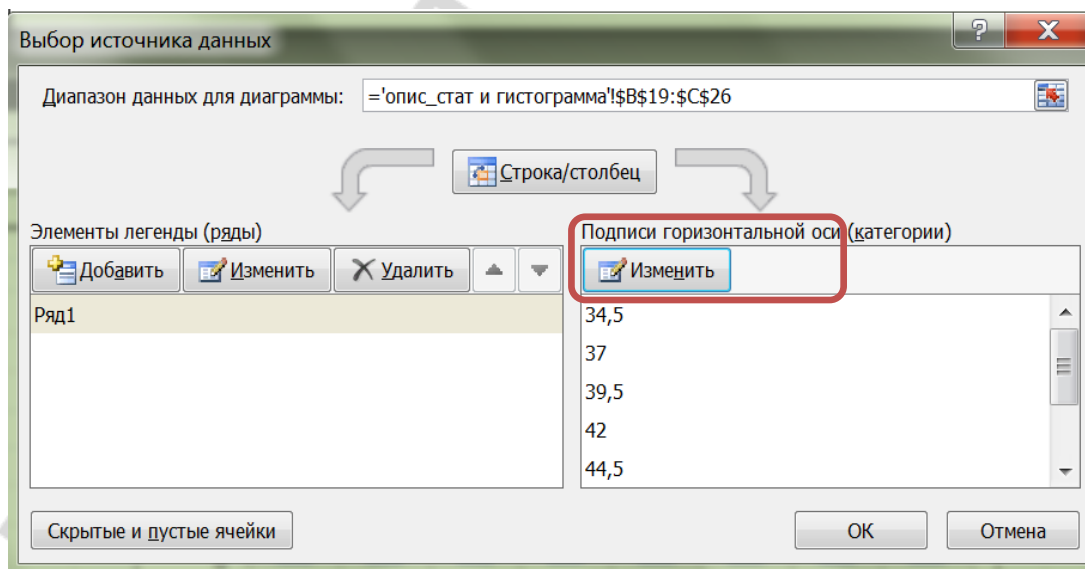


Рис. 31

- выделите мышью в таблице блок ячеек **B19:B26**, содержащий значения правых границ интервалов.

Окончательный результат показан на рис. 32.

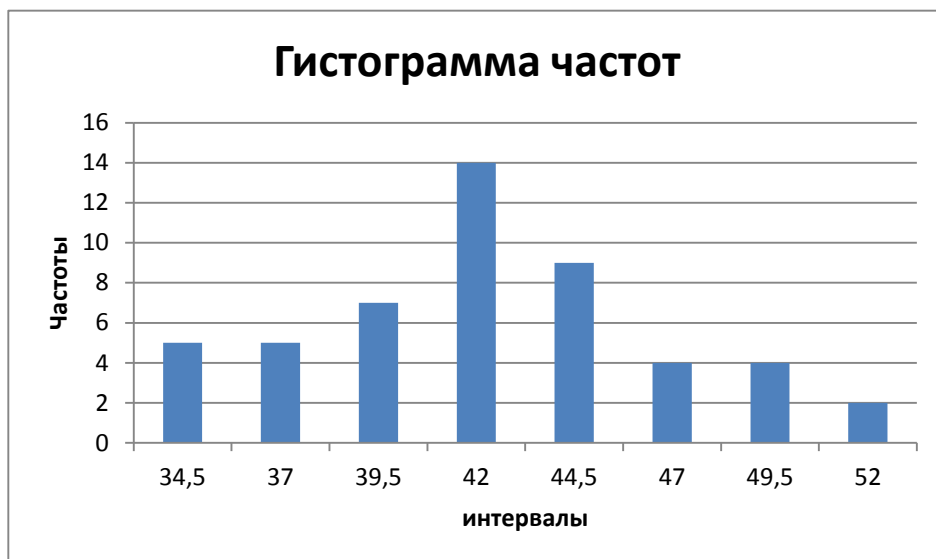


Рис. 32

11. Сохраните повторно в рабочую папку файл **Статистика.xlsx** с результатами вашей работы.

Гистограммы частот $m_i(x)$ и $m_i(x)/\Delta x$ дают хорошее наглядное представление о распределении случайной величины, но общая площадь их не нормирована к единице, а равна соответственно $n \cdot \Delta x$ и n (см. стр. 19–20). Поэтому закон распределения вероятности случайной величины X более точно представляет гистограмма относительных частот $\frac{m_i(x)}{n \cdot \Delta x}$, общая площадь столбцов которой всегда равна единице.

3.3.2. ПОСТРОЕНИЕ ГИСТОГРАММЫ ОТНОСИТЕЛЬНЫХ ЧАСТОТ

Задача. Используя данные задачи, приведенной в 3.2, постройте гистограмму относительных частот концентрации альбумина в крови женщин, откладывая по оси Y плотности относительных частот $\frac{m_i}{\Delta x \cdot n}$ для соответствующих интервалов этих величин. В данной формуле m_i — частота попадания значений величины в данный интервал, Δx — ширина интервала, n — объем выборки.

Продолжите работу в созданном ранее файле **Статистика.xlsx** с таблицей «Содержание сывороточного альбумина».

Для решения задачи выполните следующие действия:

1. Дополните таблицу частот исследуемой величины по образцу на рис. 33, выполнив следующие действия:

– введите в ячейку **D18** заголовок дополнительного столбца *Плотность относительной частоты*;

– в ячейку **D19** введите формулу для расчета плотности относительной частоты **=C19/ D16/ H7**;

– нажмите клавишу **F4**, чтобы сделать ссылки **D16** и **H7** абсолютными.

2. Скопируйте формулу из ячейки **D19** в диапазон ячеек **D20:D26**.

Результат ваших действий отображен на рис. 34.

	D
18	<u>Плотность</u> <u>относительной частоты</u>
19	=C19/\$D\$16/\$H\$7
20	=C20/\$D\$16/\$H\$7
21	=C21/\$D\$16/\$H\$7
22	=C22/\$D\$16/\$H\$7
23	=C23/\$D\$16/\$H\$7
24	=C24/\$D\$16/\$H\$7
25	=C25/\$D\$16/\$H\$7
26	=C26/\$D\$16/\$H\$7

Рис. 33

	A	B	C	D
18	№	Интервалы	Частоты	<u>Плотность</u> <u>относительной</u> <u>частоты</u>
19	1	34,5	5	0,04
20	2	37	5	0,04
21	3	39,5	7	0,06
22	4	42	14	0,11
23	5	44,5	9	0,07
24	6	47	4	0,03
25	7	49,5	4	0,03
26	8	52	2	0,02
27			50	0,4

Рис. 34

3. Желательно выполнить проверку расчетов, вычислив с помощью кнопки Автосумма  на вкладке **Главная**, в ячейках **C27** и **D27** суммы

частот и плотностей относительных частот: $\sum_{i=1}^k m_i = n$ и $\sum_{i=1}^k \frac{m_i}{n \cdot \Delta x} = \frac{1}{\Delta x}$.

	C	D
18	<u>Частоты</u>	<u>Плотность</u> <u>относительной</u> <u>частоты</u>
19	5	0,04
20	5	0,04
21	7	0,06
22	14	0,11
23	9	0,07
24	4	0,03
25	4	0,03
26	2	0,02
27	50	0,4

Рис. 35

Сумма частот должна быть равна объёму выборки — **50**, а сумма плотностей относительной частоты равна $1/2,5 = 0,4$, как показано на рис. 35. Общая площадь такой гистограммы всегда равна единице: $\Delta x \cdot \sum_{i=1}^k \frac{m_i}{n \cdot \Delta x} = 1$; и в дан-

ном случае: $2,5 \cdot 0,4 = 1$.

4. Постройте гистограмму относительных частот таким же образом, как строили предыдущую гистограмму частот, для чего:

– выделите диапазон ячеек **D19:D26**, содержащий значения плотности относительных частот;

– на вкладке **Вставка**, в группе **Диаграммы** выберите диаграмму **Гистограмма**;

– затем выберите тип диаграммы → **Гистограмма с группировкой**.

5. Используя вкладку **Работа с диаграммами/Макет**, дополните построенную гистограмму следующими элементами: **Название диаграммы** и **Названия осей** как показано на рис. 36.

6. Измените подписи горизонтальной оси — укажите на горизонтальной оси вычисленные ранее правые границы интервалов.

Результат ваших действий показан на рис. 36.

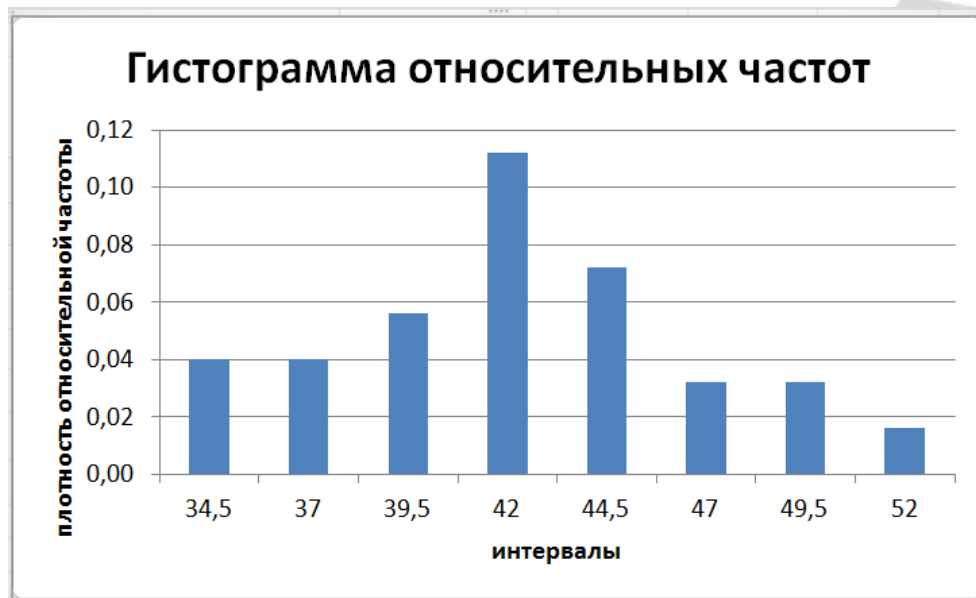


Рис. 36

3.3.3. ФОРМАТИРОВАНИЕ ГИСТОГРАММЫ

Отформатируйте первую из построенных гистограмм. Используя вкладку **Работа с диаграммами/Конструктор**, показанную на рис. 37, можно форматировать элементы диаграммы. Команды на этой вкладке задают формат выделенного в данный момент объекта (области диаграммы, области построения или столбиков гистограммы).

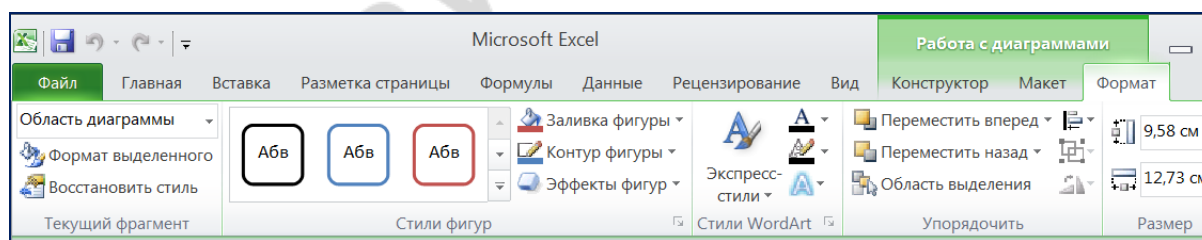


Рис. 37

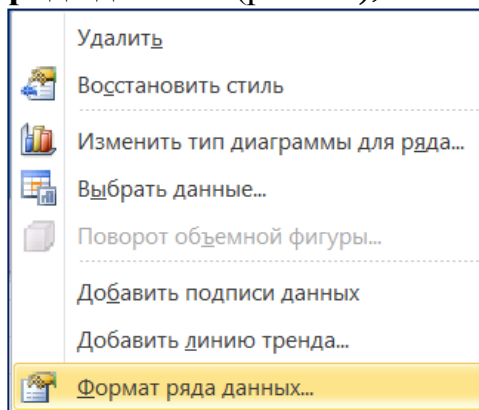
1. Измените формат *области диаграммы*, для чего:
 - выделите область диаграммы;
 - изучите возможные варианты **Стилей фигур**, выберите и примените какой-либо из стилей.
2. Уберите заливку *области построения диаграммы*:
 - выделите область построения графика,

– на вкладке **Работа с диаграммами/Формат** выберите *Заливка фигуры/Нет заливки*.

3. Измените вид *столбиков гистограммы*:

– щелчком правой кнопки мыши выделите прямоугольники гистограммы;

– выберите в появившемся контекстном меню команду **Формат ряда данных** (рис. 38);



– на вкладке **Заливка** можно выбрать в палитре другой цвет для прямоугольников гистограммы;

– на вкладке **Параметры ряда** (рис. 38) задайте *Боковой зазор «20%»*, нажмите **ОК**.

Результат ваших действий представлен на рис. 39.

Рис. 38



Рис. 39

3.3.4. РЕДАКТИРОВАНИЕ ДИАГРАММЫ

Задача. Проанализируйте влияние числа интервалов на расчет частот и поведение гистограммы. Для этого выполните процедуру расчета ширины интервала, правых границ интервалов и частот для числа интервалов 5, постройте гистограмму по новым параметрам.

1. Скопируйте фрагмент с расчетом интервалов частот ниже на рабочий лист:

- выделите диапазон ячеек **A15:D26**; примените команду **Копировать**;
- в ячейку **A33** вставьте скопированный фрагмент;
- 2. Скорректируйте скопированный фрагмент (рис. 40): в ячейке **D33** удалите старое содержимое клавишей **Del** и введите число **5**.
- 3. В ячейке **D34** отредактируйте формулу, приведите её к следующему виду **=(H6-H5)/D33**.

	A	B	C	D
32				
33		Количество интервалов		5
34		Ширина интервала		=(H6-H5)/D33

Рис. 40

- 4. В колонке **A** сократите нумерацию интервалов до **5**, удалив содержимое ячеек **A42:B44**; удалите содержимое диапазона **C37:D44**.

- 5. Измените формулы расчета границ интервалов, как показано на рис. 41:

	A	B	C	D
32				
33		Количество интервалов		5
34		Ширина интервала		=(H6-H5)/D33
35				
36	№	Интервалы	Частоты	
37	1	=H5+D34	=ЧАСТОТА(A2:E11;B37:B41)	
38	2	=B37+\$D\$34	=ЧАСТОТА(A2:E11;B37:B41)	
39	3	=B38+\$D\$34	=ЧАСТОТА(A2:E11;B37:B41)	
40	4	=B39+\$D\$34	=ЧАСТОТА(A2:E11;B37:B41)	
41	5	=B40+\$D\$34	=ЧАСТОТА(A2:E11;B37:B41)	

Рис. 41

- в ячейке **B37** замените ссылку **H23** на **H5**, нажмите **Enter**;
- в ячейке **B38** исправьте ссылку **D16** на абсолютную **D34**, нажмите **Enter**;
- скопируйте формулу из ячейки **B38** в диапазон **B39:B41**, убедитесь, что произошел перерасчет значений правых границ интервалов.
- 6. Вставьте в блок ячеек **C37:C41** функцию для расчета частот. Для этого:
 - в ячейку **C37** вставьте функцию **ЧАСТОТА** из категории **Статистические**;
 - в окне **Аргументы функции** в поле **Массив_данных** укажите диапазон ячеек **A2:E11**;
 - в поле **Массив_интервалов** укажите диапазон ячеек **B37:B41**; нажмите **ОК**;

– выделите диапазон **C37:C41**, начиная с ячейки, содержащей формулу, нажмите клавишу **F2**, а затем нажмите **Ctrl + Shift + Enter** для фиксации функции массива.

Результат вычислений показан на рис. 42.

7. Постройте новую гистограмму, состоящую из 5 столбиков.

8. Используя вкладку **Работа с диаграммами/Макет**, дополните построенную гистограмму следующими элементами: **Название диаграммы** и **Названия осей** как показано на рис. 42.

9. Измените подписи горизонтальной оси — укажите вычисленные ранее правые границы интервалов.

Результат ваших действий изображен на рис. 42.

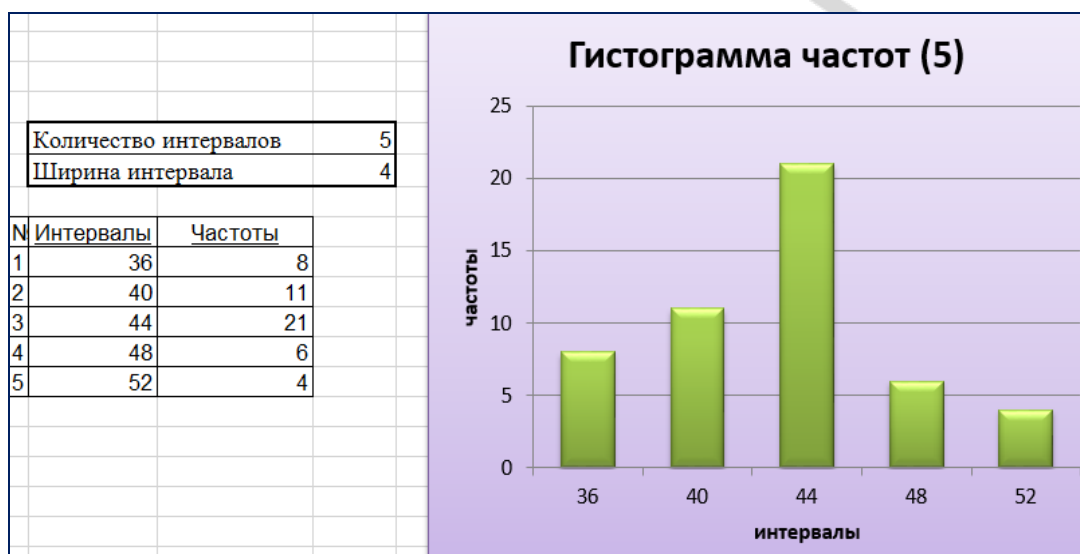


Рис. 42

Сравните две построенные диаграммы с различным количеством интервалов.

3.3.5. ПОСТРОЕНИЕ ПОЛИГОНА ЧАСТОТ

Задача. Представьте данные о концентрации сывороточного альбумина из задачи, приведенной в 3.2, в виде графика, связывающего частоту попадания в интервал с правой границей каждого интервала.

1. Постройте полигон частот, для чего выполните следующие действия:

– выделите диапазон ячеек с таблицей частот **B19:C26**;

– на вкладке **Вставка**, в группе **Диаграммы** выберите диаграмму — **Точечная**;

– затем выберите тип точечной диаграммы → **Точечная с прямыми отрезками и маркерами** (рис. 43).

Результат показан на рис. 44.

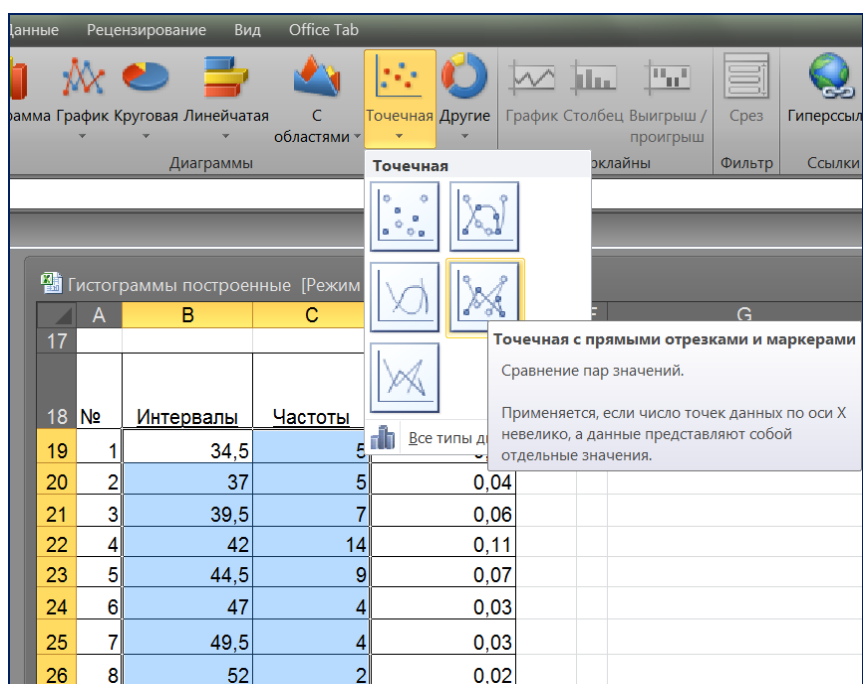


Рис. 43

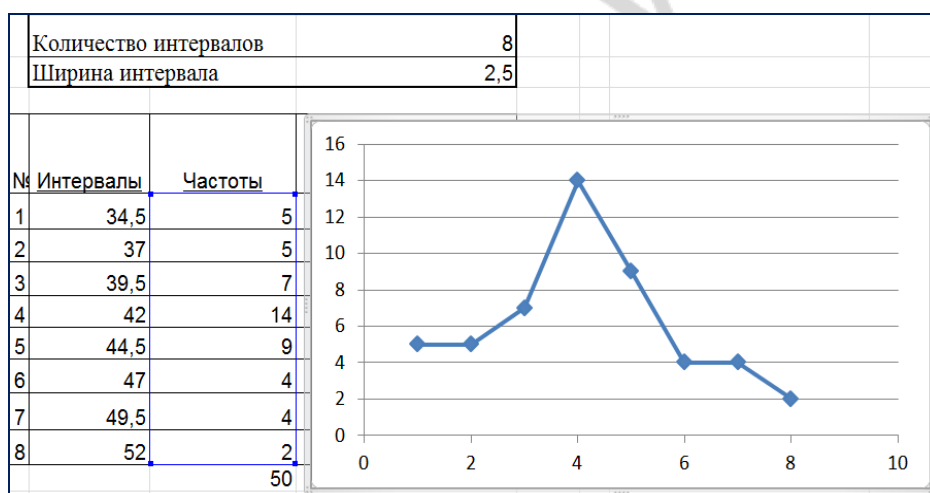


Рис. 44

2. Используя вкладку **Работа с диаграммами/Макет**, удалите *легенду* и дополните построенную гистограмму следующими элементами: **Название диаграммы** и **Названия осей** как показано на рис. 45.

3. Измените формат горизонтальной оси так, чтобы она начиналась не с нуля, а с цифры 32:

- наведите указатель мыши на горизонтальную ось, нажмите правую кнопку мыши для вызова контекстного меню (рис. 46);
- выберите в контекстном меню команду **Формат оси**;
- в открывшемся окне диалога **Формат оси** (рис. 47) на вкладке *Параметры оси* выберите следующие параметры:

– в поле минимальное значение введите «32», максимальное значение «52», цена основных делений «2», ось Y пересекает в значении «32». Нажмите **ОК**.



Рис. 45



Рис. 46

Диалоговое окно «Формат оси» с вкладкой «Параметры оси». Вкладка «Число» выделена в левом меню.

Параметры оси

минимальное значение: ☐ авто ☒ фиксированное 32,0

максимальное значение: ☐ авто ☒ фиксированное 52

цена основных делений: ☐ авто ☒ фиксированное 2

цена промежуточных делений: ☒ авто ☐ фиксированное 1,0

☐ обратный порядок значений

☐ логарифмическая шкала Основная: 10

Цена деления: нет

☐ Отображать на диаграмме

Основные: наружу

Промежуточные: нет

Подписи оси: рядом с осью

Вертикальная ось пересекает:

☐ Автовыбор

☒ Значение оси: 32,0

Рис. 47

Результат ваших действий показан на рис. 48.

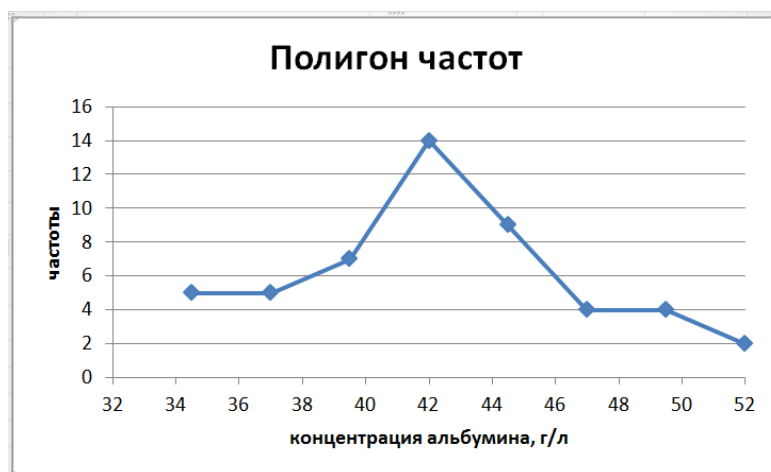


Рис. 48

Следует подчеркнуть, что результаты, получаемые с помощью встроенных статистических функций, и графики, построенные на основе этих вычислений, имеют постоянную связь с исходными данными — при изменении исходных данных результаты решения автоматически изменяются.

3.3.6. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. Анализируемый показатель — срок лечения больного при некотором заболевании. Используя приведенные в таблице данные, дополните ее параметром, необходимым для построения полигона относительных частот и постройте этот график.

Число дней лечения	17	18	20	22	23	25
Число больных с данным сроком лечения m (частота)	2	5	4	8	5	2

Задача № 2. Дан статистический ряд, содержащий сведения о длительности жизни пациентов после операции по поводу рака прямой кишки: 0, 2, 6, 6, 7, 8, 11, 11, 11, 12, 12, 12, 14, 15, 16, 17, 17, 17, 20, 24, 25, 25, 26, 26, 27, 27, 28, 30, 34, 34, 35, 38, 39, 41, 47, 54. Постройте гистограмму частот и выделите на ней другим цветом интервал наиболее вероятных значений величины.

3.4. ПРИМЕНЕНИЕ ПАКЕТА АНАЛИЗА ДЛЯ ОПРЕДЕЛЕНИЯ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ВЫБОРКИ И ПОСТРОЕНИЯ ГИСТОГРАММ

Основные вопросы:

1. Расчет выборочных характеристик случайной величины с помощью Пакета анализа.
2. Построение гистограмм с применением Пакета анализа.

3.4.1. РАСЧЕТ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ С ПОМОЩЬЮ ПАКЕТА АНАЛИЗА

В состав Microsoft Excel так же входит набор средств анализа данных (так называемый пакет анализа), который позволяет выполнить расчеты всех статистических характеристик не пошагово, как с помощью Мастера функций, а одновременно. Такой метод расчета требует **обязательного введения данных выборки в один столбец или строку.**

Задача. Используя данные задачи, приведенной в 3.2, получите значения выборочных характеристик, применив возможности Пакета анализа.

Для получения значений характеристик в этом случае выполните следующие действия:

1. Откройте файл **Статистика.xls**, созданный и сохраненный на предыдущем занятии.

2. Присвойте **Листу2** новое имя **Пакет анализа**.

3. Скопируйте с предыдущего листа диапазон ячеек **A1:E11** с данными задачи на лист **Пакет анализа** следующим образом:

– на первом листе выделите мышью диапазон **A1:E11**, примените команду **Копировать**;

– перейдите на лист **Пакет анализа**, установите курсор в ячейку **A1**, примените команду **Вставить**.

4. Данные задачи необходимо оформить в таблице в виде одного столбца, для этого:

– на листе **Пакет анализа** выделите мышью диапазон **B2:B11**, примените команду **Вырезать**, вставьте вырезанный фрагмент в ячейку **A12**;

– далее выделите и вырежьте диапазон **C2:C11**, вставьте вырезанный фрагмент в ячейку **A22**;

– описанным образом вырежьте по очереди и скопируйте диапазоны с данными **D2:D11**, **E2:E11**

5. Запустите пакет анализа:

– выберите на вкладке **Данные** команду **Анализ данных** (рис. 49);

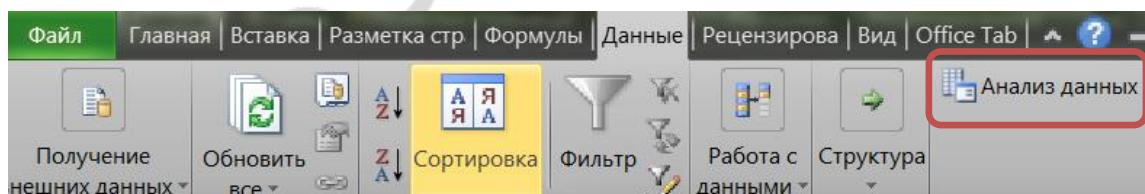


Рис. 49

– если эта кнопка на вкладке отсутствует, чтобы добавить ее, зайдите в меню **Файл/ Параметры**; в окне диалога выберите **Надстройки**, там нажмите кнопку **Перейти**; в открывшемся окне выделите мышью **Пакет анализа**; нажмите **ОК** (рис. 50); теперь на вкладке **Данные** появится кнопка **Анализ данных**, щелкните по ней.

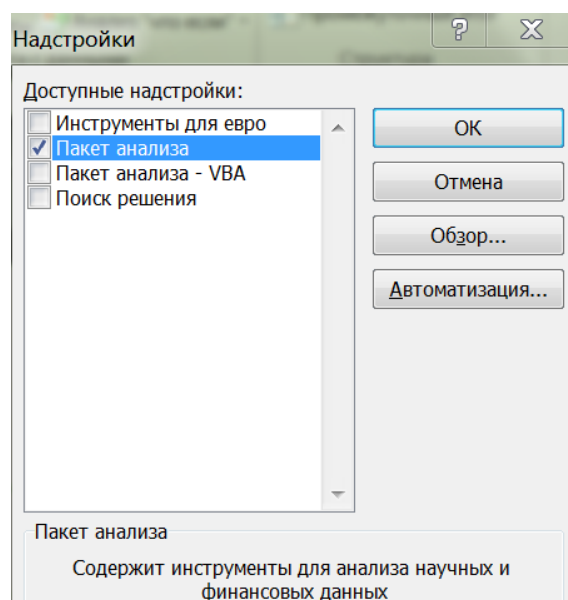


Рис. 50

6. В открывшемся окне (рис. 51) выделите щелчком необходимый инструмент анализа — **Описательная статистика**, подтвердите выбор, нажав **ОК**. В результате этих действий на экране появится окно диалога **Описательная статистика**.

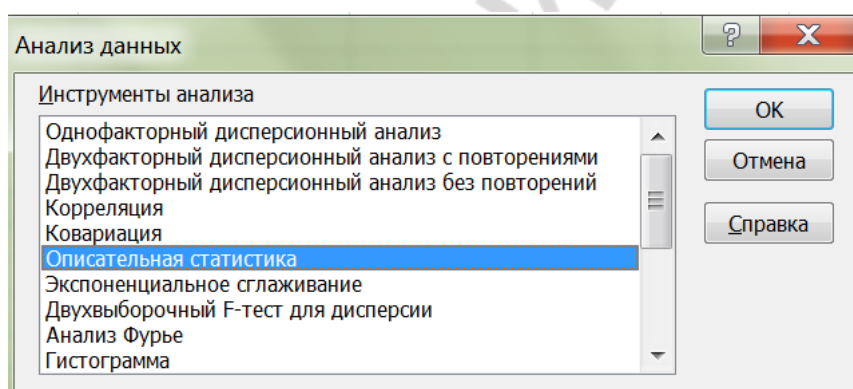


Рис. 51

7. Укажите в окне **Описательная статистика** (рис. 52) следующие параметры:

- **входной интервал** — адреса ячеек, содержащие анализируемые данные (в нашем примере это **A2:A51**);
- **выходной интервал** — щелкните мышью на ячейку в электронной таблице, начиная с которой будут выводиться результаты анализа, в нашем примере это ячейка **F2**;
- в разделе **Группирование** убедитесь, что переключатель установлен в положение «по столбцам», так как данные введены в столбец;
- щелчком мыши установите флажок **Итоговая статистика** для вывода результатов анализа; подтвердите ввод параметров, нажав **ОК**.

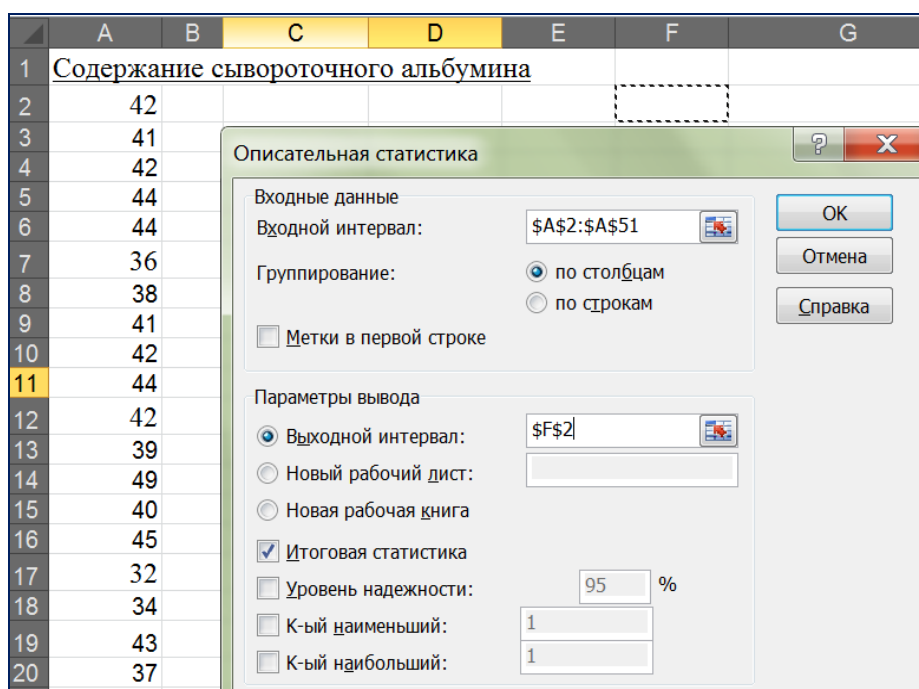


Рис. 52

Результаты расчетов приведены на рис. 53.

	A	B	C	D	E	F	G
1	Содержание сывороточного альбумина						
2	42					Столбец1	
3	41						
4	42				Среднее		41,38
5	44				Стандартная ошибка		0,67
6	44				Медиана		42
7	36				Мода		42
8	38				Стандартное отклонение		4,75
9	41				Дисперсия выборки		22,53
10	42				Эксцесс		-0,18
11	44				Асимметричность		0,03
12	42				Интервал		20
13	39				Минимум		32
14	49				Максимум		52
15	40				Сумма		2069
16	45				Счет		50

Рис. 53

Сопоставьте результаты, полученные с помощью Пакета анализа на листе **Пакет анализа**, с результатами расчетов выполненных с помощью функций на листе **Описательная статистика** (рис. 21, б и 53).

3.4.2. ПОСТРОЕНИЕ ГИСТОГРАММЫ С ПОМОЩЬЮ ПАКЕТА АНАЛИЗА

Постройте на листе **Пакет анализа**, гистограмму, выполнив следующие действия.

1. Запустите пакет анализа: выберите на вкладке **Данные** команду **Анализ данных**.

2. В открывшемся окне (рис. 54) выделите щелчком **Гистограмма** и подтвердите выбор, нажав **ОК**. В результате этих действий на экране появится диалоговое окно **Гистограмма**.

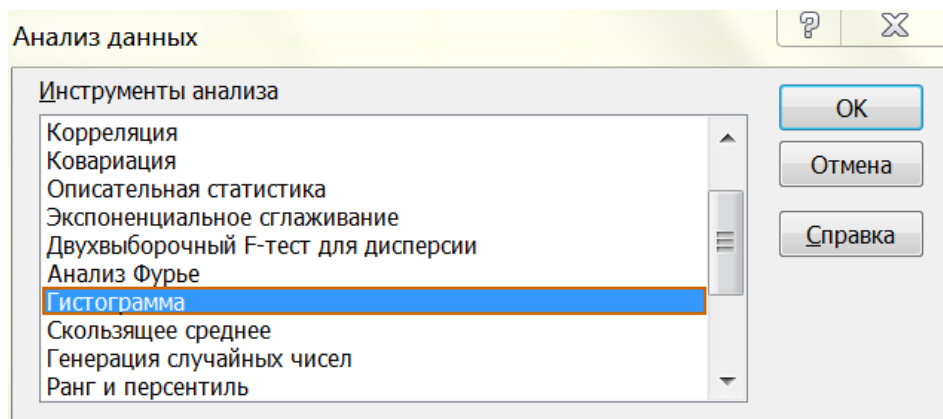


Рис. 54

3. Укажите в окне **Гистограмма** (рис. 55) следующие параметры:

- **входной интервал** — адреса ячеек, содержащие анализируемые данные. Для этого выделите мышью блок ячеек с данными (в нашем примере это **A2:A51**);
- **выходной интервал** — щелкните мышью на ячейку в электронной таблице, начиная с которой будут выводиться результаты анализа, в нашем примере это ячейка **F19**;

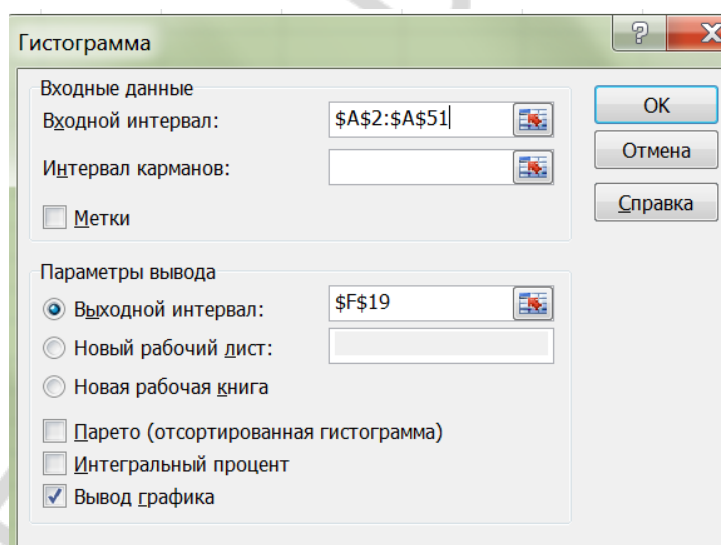


Рис. 55

4. Установите флажок **Вывод графика** для автоматического создания диаграммы, подтвердите параметры, нажав **ОК**.

Результат отображен на рис. 56.

Сравните гистограмму, построенную вручную на листе **Описательная статистика** (см. рис. 32), и гистограмму, построенную с помощью Пакета анализа, на листе **Пакет анализа** (рис. 56).

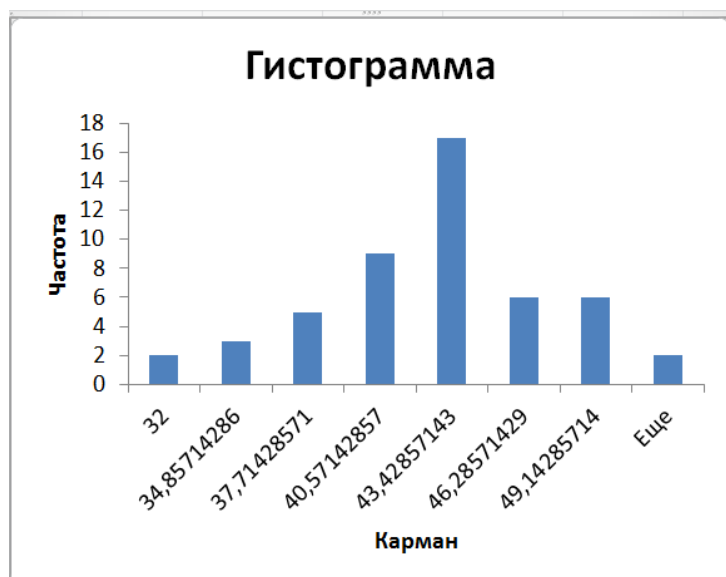


Рис. 56

Отметим также, что результаты, полученные с помощью Пакета анализа, не имеют постоянной связи с исходными данными. При изменении исходных данных необходимо повторно выполнить анализ.

3.4.3. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача. Анализируемый показатель — значение гематокрита (Hct) у больных в критическом состоянии. При поступлении в стационар получены приведенные ниже данные (%).

34	41	46	28	39	30	25	41	42	31	25	20	26	28	27	37	44	41	32
30	38	20	25	31	33	36	32	36	43	31	27	30	38	35	42	33	21	34

Вычислите следующие статистические характеристики для этого показателя: выборочные среднее, медиану, моду, минимальное и максимальное значения, дисперсию, стандартное отклонение, объем выборки, коэффициент асимметрии и эксцесс. Постройте гистограмму частот. Можно ли утверждать по этим выборочным сведениям, что данный показатель генеральной совокупности распределен по нормальному закону?

3.5. ИНТЕРВАЛЬНАЯ ОЦЕНКА ПАРАМЕТРОВ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ПО ЕЕ ВЫБОРКЕ. РАСЧЕТ ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ

Основные вопросы:

1. Расчет доверительных интервалов для генерального среднего.
2. Расчет границ доверительного интервала для генерального среднего.

Задача. Используя данные задачи, приведенной в подразд. 3.2, постройте доверительные интервалы для среднего, считая, что этот показатель в генеральной совокупности распределен по нормальному закону.

Доверительную вероятность принять $\gamma = 0,95 = 95 \%$, следовательно, уровень значимости $\alpha = 0,05$. Используйте уже полученные значения объема (размера) выборки, выборочного среднего, стандартного отклонения.

3.5.1. ПОДГОТОВКА ТАБЛИЦЫ ДАННЫХ

Для подготовки таблицы данных выполните следующие действия:

1. Откройте сохраненную на предыдущих занятиях книгу под именем **Статистика.xlsx** в своей рабочей папке.
2. Присвойте чистому листу **Лист3** новое имя **Доверит интервалы**.
3. Скопируйте таблицу с исходными данными задачи «Содержание сывороточного альбумина» и вычисленными ранее числовыми характеристиками с листа **Описательная статистика** на новый лист **Доверит интервалы** (рис. 57).

	A	B	C	D	E	F	G	H
1	Содержание сывороточного ал					Описательная статистика		
2	42	41	42	44	44	Среднее		41,38
3	36	38	41	42	44	Мода		42
4	42	39	49	40	45	Медиана		42
5	32	34	43	37	39	Минимум		32
6	41	39	48	42	43	Максимум		52
7	33	43	35	32	34	Объем выборки		50
8	39	35	43	44	47	Асимметрия		0,03
9	40	39	42	41	46	Экссесс		-0,18
10	37	49	41	39	43	Дисперсия		22,53
11	42	47	48	51	52	Стандартное отклонение		4,75
12						Вариационный размах		20
13						Стандартная ошибка		0,67

Рис. 57

3.5.2. РЕДАКТИРОВАНИЕ ТАБЛИЦЫ ДАННЫХ

Для расчета доверительного интервала потребуются следующие характеристики: среднее, объем выборки, дисперсия, стандартное отклонение.

1. На данном листе в скопированном блоке удалите ячейки с расчетами, невостробованными в данной задаче (выделенные пунктиром на рис. 57 диапазоны **G3:H6**; **G8:H9** и **G12**):

- выделите мышью диапазон ячеек **G3:H6**; в контекстном меню выберите команду **Удалить**;
- в окне диалога *Удаление ячеек* (рис. 58) поставьте переключатель в позицию *ячейки со сдвигом вверх*.

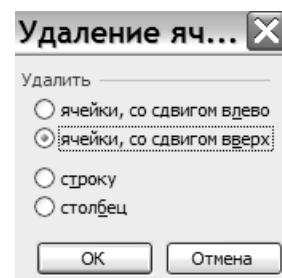


Рис. 58

2. Описанным выше способом удалите остальные ненужные формулы по образцу на рис. 59.

3. Дополните таблицу данных — введите в ячейки **G7** и **G8** названия величин (*Уровень значимости* и *Доверительный интервал*), а в **H7** введите числовое значение (0,05) как показано на рис. 60:

	A	B	C	D	E	F	G	H
1	Содержание					Описательная статистика		
2	42	41	42	44	44	Среднее		41,38
3	36	38	41	42	44	Объем выборки		50
4	42	39	49	40	45	Дисперсия		22,53
5	32	34	43	37	39	Стандартное отклонение		4,75
6	41	39	48	42	43	Стандартная ошибка		0,67
7	33	43	35	32	34			
8	39	35	43	44	47			
9	40	39	42	41	46			
10	37	49	41	39	43			
11	42	47	48	51	52			

Рис. 59

	A	B	C	D	E	F	G	H
1	Содержание					Описательная статистика		
2	42	41	42	44	44	Среднее		41,38
3	36	38	41	42	44	Объем выборки		50
4	42	39	49	40	45	Дисперсия		22,53
5	32	34	43	37	39	Стандартное отклонение		4,75
6	41	39	48	42	43	Стандартная ошибка		0,67
7	33	43	35	32	34	Уровень значимости		0,05
8	39	35	43	44	47	Доверительный интервал		
9	40	39	42	41	46			
10	37	49	41	39	43			
11	42	47	48	51	52			

Рис. 60

3.5.3. РАСЧЕТ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ДЛЯ ГЕНЕРАЛЬНОГО СРЕДНЕГО

1. Для расчета доверительного интервала в ячейку **H8** вставьте функцию **ДОВЕРИТ.НОРМ** из категории **Статистические**.

2. В окне **Аргументы функции** (рис. 61) укажите адреса ячеек, содержащих предварительно рассчитанные стандартное отклонение, объем (размер) выборки и заданный уровень значимости:

Аргументы функции

ДОВЕРИТ.НОРМ

Альфа = 0,05

Станд_откл = 4,746169239

Размер = 50

= 1,31554682

Возвращает доверительный интервал для среднего генеральной совокупности с использованием нормального распределения.

Размер размер выборки.

Рис. 61

– в поле **Альфа** введите уровень значимости **0,05** или адрес ячейки **H7**;
– в поле **Станд_откл** укажите **H5** — адрес ячейки, содержащей стандартное отклонение;

Уровень значимости	0,05
Доверительный интервал	1,32

Рис. 62

– в поле **Размер** укажите ячейку **H3**, содержащую вычисленный объем выборки; нажмите **ОК**.

Полученный числовой результат отобразится в ячейке **H8**, как показано на рис. 62.

3.5.4. РАСЧЕТ ГРАНИЦ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ДЛЯ ГЕНЕРАЛЬНОГО СРЕДНЕГО ЗНАЧЕНИЯ

1. Заполните диапазон **J1:J3** текстовыми значениями по образцу, отображенному на рис. 63.

	Г	Н	И	Ж	К
1	Описательная статистика			Доверительный интервал для среднего	
2	Среднее	41,38		нижняя граница	
3	Объем выборки	50		верхняя граница	
4	Дисперсия	22,5261			
5	Стандартное отклонение	4,74617			
6	Стандартная ошибка	0,67121			
7	уровень значимости	0,05			
8	Доверительный интервал	1,31555			

Рис. 63

2. Увеличьте ширину столбца **Ж**, дважды щелкнув правую границу заголовка столбца **Ж**.

3. В ячейку **К2** введите формулу **=Н2-Н8** для расчета нижней границы доверительного интервала:

4. Введите в ячейку **К3** формулу **=Н2+Н8** для расчета верхней границы доверительного интервала:

Ж	К
Доверительный интервал для среднего	
нижняя граница	40,06
верхняя граница	42,70

Вычисленные границы доверительного интервала для среднего отображены на рис. 64.

Рис. 64

Значение сывороточного альбумина с вероятностью **95 %** лежит в интервале $40,06 < x < 42,70$.

Задание.

Самостоятельно вычислите в столбце **Л** границы нового доверительного интервала, приведите результат в виде $M_t(X) = \bar{x}_v \pm m$ (где m — стандартная ошибка, при $t = 1$) и убедитесь, что границы доверительного интервала сужаются $40,71 < x < 42,05$. Приведенный интервал будет покрывать истинное значение величины с вероятностью **68 %**.

	А	В	С	Д	Е	Г	Н	И	Ж	К	Л
1	Содержание сывороточного					Описательная статистика			Доверительный интервал для среднего		
2	42	41	42	44	44	Среднее	41,38		нижняя граница	40,06	40,71
3	36	38	41	42	44	Объем выборки	50		верхняя граница	42,70	42,05
4	42	39	49	40	45	Дисперсия	22,53			95%	68%
5	32	34	43	37	39	Стандартное отклонение	4,75				
6	41	39	48	42	43	Стандартная ошибка	0,67				
7	33	43	35	32	34	Уровень значимости	0,05				
8	39	35	43	44	47	Доверительный интервал	1,32				
9	40	39	42	41	46						
10	37	49	41	39	43						
11	42	47	48	51	52						

Рис. 65

3.5.5. РЕДАКТИРОВАНИЕ ТАБЛИЦЫ РАСЧЕТА ДОВЕРИТЕЛЬНЫХ ИНТЕРВАЛОВ. АНАЛИЗ ЗАВИСИМОСТИ ГРАНИЦ ИНТЕРВАЛОВ ОТ ОБЪЕМА ВЫБОРКИ

Задача. Используя данные задачи, приведенной в 3.2, и таблицу расчета доверительных интервалов, проанализируйте, как изменятся границы доверительных интервалов при уменьшении объема выборки до 35.

1. Вставьте в книгу **Статистика.xlsx** новый рабочий лист.
2. Переименуйте новый лист, дав ему имя **Дов инт 35**.
3. Скопируйте на новый лист **Дов инт 35** условие и вычисления задачи, приведенной в подразд. 3.5.6 (см. рис. 65).
4. Уменьшите объем выборки с 50 до 35, удалив данные из ячеек **A9:E11** как показано на рис. 66.
5. Пользуясь строкой формул, отредактируйте скопированные формулы для расчета числовых характеристик выборки, находящиеся в ячейках **H2:H5**, изменив в них адрес диапазона ячеек с **A2:E11** на **A2:E8**, где находятся данные для вычислений.
6. Проследите, как во всех остальных формулах для расчета доверительных интервалов автоматически выполняется перерасчет.

Результат ваших действий отображен на рис. 66.

	A	B	C	D	E	F	G	H	I	J	K
	<u>Содержание</u>									<u>Доверительный интервал</u>	
1	сывороточного						<u>Описательная статистика</u>			<u>для среднего</u>	
2	42	41	42	44	44		Среднее	40,34		нижняя граница	38,87
3	36	38	41	42	44		Объем выборки	35		верхняя граница	41,81
4	42	39	49	40	45		Дисперсия	19,70			
5	32	34	43	37	39		Стандартное отклонение	4,44			
6	41	39	48	42	43		Стандартная ошибка	0,75			
7	33	43	35	32	34		Уровень значимости	0,05			
8	39	35	43	44	47		Доверительный интервал	1,47			
9											
10											
11											

Рис. 66

Среднее значения для сывороточного альбумина в этом случае с вероятностью 95 % лежит в интервале $38,87 < \sigma < 41,81$.

3.6. ПРОВЕРКА ПРИНАДЛЕЖНОСТИ РАСПРЕДЕЛЕНИЯ ВЫБОРКИ К ТЕОРЕТИЧЕСКОМУ НОРМАЛЬНОМУ

Задача. Используя данные задачи, приведенной в подразд. 3.2, установите соответствие выборочного распределения теоретическому нормальному следующим образом: рассчитайте среднее и стандартное отклонение выборочных значений, абсолютные значения отклонений вы-

борочных значений от среднего, а затем проверьте выполнение следующих условий:

- 99,7 % отклонений от среднего меньше $3S$;
- 95,5 % отклонений от среднего меньше $2S$;
- 68,3 % отклонений от среднего меньше S , где S — стандартное отклонение выборочного среднего.

Рассмотрим последовательность необходимых действий.

Вставьте в книгу под именем **Статистика** новый лист, присвойте этому листу имя **Проверка на норм.**

Скопируйте на этот лист таблицу с исходными данными задачи «Содержание сывороточного альбумина» — диапазон ячеек **A1:E11**.

Введите в ячейки **G13:G20** следующие названия вычисляемых характеристик по образцу на рис. 67.

	G
13	среднее
14	объём выборки
15	значение S
16	значение 2S
17	значение 3S
18	68,3%
19	95,5%
20	99,7%

Рис. 67

Рассчитайте в ячейке **H13** среднее значение, используя известную вам функцию **СРЗНАЧ** для диапазона ячеек с данными **A2:E11** (рис. 68).

Вычислите в ячейке **H14** объём выборки, используя функцию **СЧЁТ**, а в ячейке **H15** стандартное отклонение, используя функцию **СТАНДОТКЛОН.В**, каждый раз указывая диапазон ячеек с данными **A2:E11**.

	G	H
13	среднее	=СРЗНАЧ(A2:E11)
14	объём выборки	=СЧЁТ(A2:E11)
15	значение S	=СТАНДОТКЛОН(A2:E11)
16	значение 2S	=2*H15
17	значение 3S	=3*H15

Рис. 68

В ячейке **H16** вычислите значение $2S$, используя формулу **=2*H15**, аналогично в ячейке **H17** вставьте формулу **=3*H15** для вычисления значения $3S$.

Введите в ячейку **A13** текст «отклонения от среднего».

	A	B
13	отклонения от среднего	
14	=ABS(A2-\$H\$13)	

Рис. 69

Рассчитайте массив отклонений выборочных значений от среднего, для чего в ячейку **A14** введите функцию **=ABS(A2-\$H\$13)** как показано на рис. 69 (функция **ABS** из категории *Математические*).

– чтобы сделать ссылку **H13** абсолютной, нажмите функциональную клавишу **F4**; подтвердите, нажав **ОК**.

– скопируйте введенную в ячейку **A14** формулу вниз до **A23**, затем вправо до **E23**.

Результат расчетов отображен на рис. 70.

	A	B	C	D	E
13	отклонения от среднего				
14	0,62	0,38	0,62	2,62	2,62
15	5,38	3,38	0,38	0,62	2,62
16	0,62	2,38	7,62	1,38	3,62
17	9,38	7,38	1,62	4,38	2,38
18	0,38	2,38	6,62	0,62	1,62
19	8,38	1,62	6,38	9,38	7,38
20	2,38	6,38	1,62	2,62	5,62
21	1,38	2,38	0,62	0,38	4,62
22	4,38	7,62	0,38	2,38	1,62
23	0,62	5,62	6,62	9,62	10,6

Рис. 70

Подсчитайте в ячейках **H18:H20** сколько отклонений составляет 68,3; 95,5 и 99,7 % от их общего количества (рис. 71):

	G	H
18	68,3%	=0,683*N14
19	95,5%	=0,955*N14
20	99,7%	=0,997*N14

Рис. 71

– в ячейку **H18** вставьте формулу: **=0,683*N14** (где N14 — объем выборки).

– таким же образом вычислите 95,5 и 99,7 % от общего количества в ячейках **H19** и **H20** соответственно, как показано на рис. 71.

Результат вычислений отображен на рис. 72.

	G	H
13	среднее	41,38
14	объем	50
15	значение S	4,75
16	значение 2S	9,49
17	значение 3S	14,24
18	68,3%	34,15
19	95,5%	47,75
20	99,7%	49,85

Рис. 72

	G	H	I
21	Условие	Выполнение условия	
22	< S		
23	< 2S		
24	< 3S		

Рис. 73

Создайте таблицу условий проверки распределения на нормальность (рис. 73), для чего:

– введите в ячейки **G21:G24** текстовые значения по образцу на рис. 73;

– введите в ячейки **H22:H24** функции для вычисления критерия выполнения условия (подсчет отклонений удовлетворяющих заданному условию) как на рис. 74, используя функцию **СЧЕТЕСЛИ** из Категории **Статистические**;

– в окне **Аргументы функции** (рис. 75) в поле *Диапазон* укажите ссылку на диапазон ячеек **A14:E23**;

	G	H
21	Условие	Выполнение условия
22	68,3% < S	=СЧЕТЕСЛИ(A14:E23;"<4,75")
23	95,5% < 2S	=СЧЕТЕСЛИ(A14:E23;"<9,49")
24	99,7% < 3S	=СЧЕТЕСЛИ(A14:E23;"<14,24")

Рис. 74

Аргументы функции

СЧЕТЕСЛИ

Диапазон: A14:E23 = {0,619999999999999}

Критерий: "<4,75" = "<4,75"

= 34

Подсчитывает количество непустых ячеек в диапазоне, удовлетворяющих заданному условию.

Диапазон: диапазон, в котором подсчитывается количество непустых ячеек

Рис. 75

- в поле *Критерий* введите <4,75; нажмите **ОК**.
- аналогичным образом введите формулы в ячейках **H23** и **H24** по образцу на рис. 74.

Результат ваших действий отображен на рис. 76.

	A	B	C	D	E	F	G	H	I
13	отклонение от среднего						среднее	41,38	
14	0,62	0,38	0,62	2,62	2,6		объем выборки	50	
15	5,38	3,38	0,38	0,62	2,6		значение S	4,75	
16	0,62	2,38	7,62	1,38	3,6		значение 2S	9,49	
17	9,38	7,38	1,62	4,38	2,4		значение 3S	14,24	
18	0,38	2,38	6,62	0,62	1,6		68,3%	34,15	
19	8,38	1,62	6,38	9,38	7,4		95,5%	47,75	
20	2,38	6,38	1,62	2,62	5,6		99,7%	49,85	
21	1,38	2,38	0,62	0,38	4,6		<u>Условие</u>	<u>Выполнение условия</u>	
22	4,38	7,62	0,38	2,38	1,6		68,3%<S	34	
23	0,62	5,62	6,62	9,62	11		95,5%<2S	48	
24							99,7%<3S	50	

Рис. 76

Так как число отклонений от среднего должно быть целым, округлив значения в ячейках **H18:H20** до целого, убедитесь, что выполняются все 3 условия:

- 99,7 % отклонений от среднего меньше 3S;
- 95,5 % отклонений от среднего меньше 2S;
- 68,3 % отклонений от среднего меньше S.

Следовательно, распределение выборки соответствует нормальному.

3.7. ПРОВЕРКА ГИПОТЕЗ, СВЯЗАННЫХ С ПАРАМЕТРАМИ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Выбор необходимого метода для решения задачи о сравнении параметров распределения выборок представлен ниже в таблице:

Формулировка задачи в прикладной постановке	Формулировка задачи в статистической постановке	Дополнительные условия		Применяемый метод
Сравнение показателей контрольной и экспериментальной выборок	Проверка гипотезы о равенстве средних (центров распределения) в двух независимых выборках	Нормальный закон распределения	Дисперсии выборок равны	t-критерий (Стьюдента) при равных дисперсиях
			Дисперсии выборок не равны	t-критерий (Стьюдента) при неравных дисперсиях

Формулировка задачи в прикладной постановке	Формулировка задачи в статистической постановке	Дополнительные условия		Применяемый метод
			Без предположений о дисперсиях (но при одинаковом размере выборок)	t-критерий (Стьюдента) без предположений о дисперсиях
		Закон распределения отличный от нормального, или данные измеряются в дискретной шкале	Дисперсии выборок равны	Манна–Уитни (U-критерий Уилкоксона–Манна–Уитни)
			Без предположений о дисперсиях	Двухвыборочный критерий Уилкоксона, медианный
Сравнение показателей выборки до и после эксперимента	Проверка гипотезы о равенстве средних в двух зависимых выборках	Нормальный закон распределения		t-критерий (Стьюдента) для связанных выборок
		Закон распределения отличный от нормального или данные измеряются в дискретной шкале		Знаковый, одновыборочный критерий Уилкоксона
Можно ли считать, что среднее значение показателя равно некоторому номинальному значению?	Проверка гипотезы о равенстве среднего константе	Нормальный закон распределения		t-критерий (Стьюдента)
		Закон распределения, отличный от нормального или данные измеряются в дискретной шкале		Критерий Гупта, знаковый
Сравнение рассеяния исследуемого показателя в двух выборках	Проверка гипотезы о равенстве дисперсий двух выборок (или их принадлежности к одной генеральной совокупности)	Нормальный закон распределения		F-критерий (Фишера–Снедекора)

Для выборок с нормальным законом распределения рассмотрим подробнее следующие вопросы:

Основные вопросы:

1. Проверка гипотезы о значении математического ожидания нормальной случайной величины (одновыборочный t-критерий Стьюдента¹).

¹ В табличном процессоре Excel термин критерий заменяется словом тест.

2. Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей при неизвестных, но одинаковых дисперсиях (двухвыборочный t-критерий Стьюдента).

3. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей (F-критерий Фишера–Снедекора).

4. Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными дисперсиями, зависимые выборки (парный двухвыборочный t-критерий Стьюдента).

3.7.1. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ ЗАДАННОМУ ЧИСЛОВОМУ ЗНАЧЕНИЮ

Задача № 1. Случайная величина X — сердечный индекс (л/мин·м²). Известно, что в популяции здоровых людей его среднее значение $M(X) = A = 3,5$ л/мин·м². По данным выборки из 112 пациентов, находящихся в критическом состоянии, среднее выборочное значение $\bar{x}_в = 2,45(\frac{\text{л}}{\text{мин} \cdot \text{м}^2})$, стандартное отклонение равно 1,32 л/мин·м². Известно, что у критически больных пациентов кровообращение замедленно. Нужно установить, будет ли у критических больных:

- $M_1(X)$ не равно 3,5 л/мин·м²;
- $M_1(X)$ меньше чем 3,5 л/мин·м².

Выбираем нулевую гипотезу $H_0: M_1(X) = A = 3,5$ л/мин·м².

Рассмотрим по отдельности две разные конкурирующие гипотезы:

- $H_1: M_1(X) \neq A = 3,5$ л/мин·м² (двусторонняя критическая область);
- $H_2: M_2(X) < A = 3,5$ л/мин·м² (левосторонняя критическая область).

Уровень значимости примем $\alpha = 0,05$ в обоих случаях.

Решение.

1. Создайте новый файл и сохраните его под именем «Гипотезы» в рабочую папку.

2. Присвойте первому листу имя «О равенстве А».

3. Введите данные задачи № 1 согласно образцу на рис. 77.

4. Выполните в ячейке В7 расчет наблюдаемой t-статистики (рис.84а) по

$$t = \frac{(\bar{X}_в - A)\sqrt{n}}{S}$$

формуле
5. Присвойте первому листу имя «О равенстве А».

	А	В	С
1	Сердечный индекс	$\neq A$	$\leq A$
2	предполагаемое значение	3,5	3,5
3	объем выборки	112	112
4	среднее выборочное	2,45	2,45
5	стандартное отклонение	1,32	1,32
6			
7	t наблюдаемое		
8	t критическое двустороннее		
9	t критическое одностороннее		

Рис. 77

6. Так как проверять будем отдельно две конкурирующие гипотезы а) и б), то скопируйте данные из диапазона В2:В7 в ячейки С2:С7 (для расчета случая б)), затем в ячейки В8 и С9 вставьте функции для вычисления дву- и одностороннего t критического согласно рис. 78.

	А	В		А	В
7	t наблюдаемое	$=(B4-B2)*КОРЕНЬ(B3)/B5$	7	t наблюдаемое	-8,42
	<i>а</i>			<i>б</i>	

Рис. 78: *а* — в режиме отображения формул; *б* — в режиме результатов расчётов

7. Чтобы проверить первую альтернативную гипотезу $H_1: M_1(X) \neq A$ вычислите в ячейке В8 значение t -критическое двустороннее, для этого вставьте в ячейку В8 статистическую функцию СТЬЮДЕНТ.ОБР.2Х (так как критическая область двусторонняя);

- в окне диалога Аргументы функции (рис. 79) в поле Вероятность введите значение принятого нами уровня значимости — 0,05¹;
- в поле Степени_свободы введите число 111 (объем выборки минус один: $n - 1 = 111$).

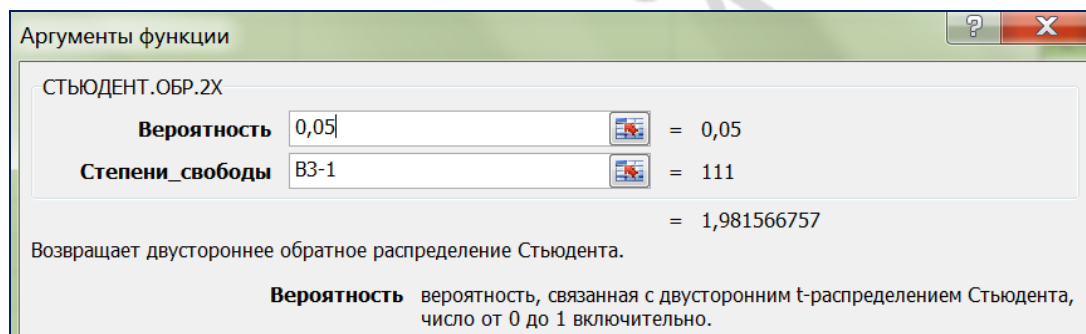


Рис. 79

Результат вычислений отображен на рис. 81.

8. Чтобы проверить вторую альтернативную гипотезу $H_2: M_2(X) < A$, вычислите в ячейке С9 значение t -критическое левостороннее, для чего:

- вставьте в ячейку С9 статистическую функцию СТЬЮДЕНТ.ОБР (так как критическая область односторонняя);
- в окне диалога Аргументы функции (рис. 80) в поле Вероятность² введите $(1-\alpha)$ значение заданного уровня значимости, здесь это $(1-0,05)$;
- в поле Степени_свободы введите число 111, (объем выборки минус один).

¹ В поле Вероятность вводят значимость, соответствующую двустороннему распределению Стьюдента. Например, для уровня значимости 0,05 и числа степеней свободы 10 двустороннее значение вычисляется СТЬЮДЕНТ.ОБР.2Х(0,05;10).

² Одностороннее t -значение может быть получено при замене аргумента «вероятность» на $(1 - \text{значимость})$. Поэтому одностороннее значение для той же значимости и того же числа степеней свободы может быть вычислено формулой СТЬЮДЕНТ.ОБР((1-0,05);10).

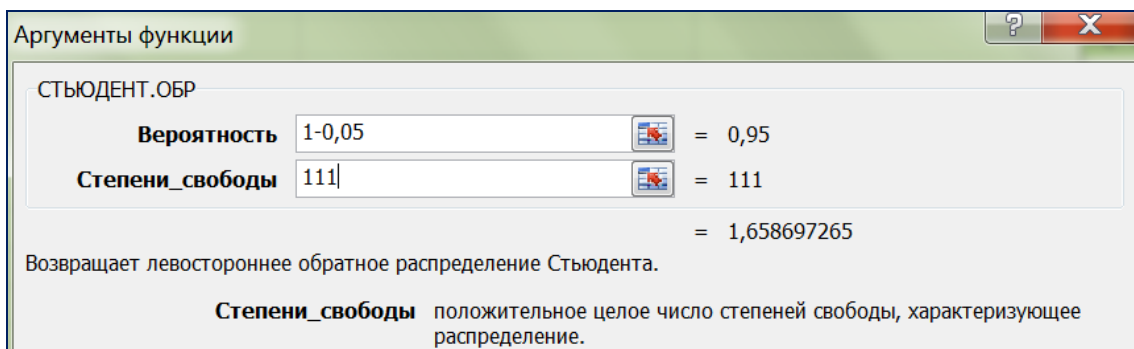


Рис. 80

Результат вычислений отображен на рис. 81, б.

	А	В	С
1	Сердечный индекс	$\neq A$	$\leq A$
2	предполагаемое значение	3,5	3,5
3	объем выборки	112	112
4	среднее выборочное	2,45	2,45
5	стандартное отклонение	1,32	1,32
6			
7	t наблюдаемое	= (B4-B2)*КОРЕНЬ(B3)/B5	= (C4-C2)*КОРЕНЬ(C3)/C5
8	t критическое двустороннее	=СТЮДЕНТ.ОБР.2Х(0,05;B3-1)	
9	t критическое одностороннее		=СТЮДЕНТ.ОБР(1-0,05;111)
10			=СТЮДРАСПОБР(0,05*2;111)

а

	А	В	С
1	Сердечный индекс	$\neq A$	$\leq A$
2	предполагаемое значение	3,5	3,5
3	объем выборки	112	112
4	среднее выборочное	2,45	2,45
5	стандартное отклонение	1,32	1,32
6			
7	t наблюдаемое	-8,42	-8,42
8	t критическое двустороннее	1,98	
9	t критическое одностороннее		1,66

б

Рис. 81: а — в режиме отображения формул; б — в режиме результатов расчётов

Для расчета значения **t**-критического можно использовать ещё одну статистическую функцию **СТЮДРАСПОБР**¹, которую оставили для совместимости с Excel 2007 и более ранних версий.

¹ В окне диалога в поле *Вероятность* вводят значимость, соответствующую двустороннему распределению Стьюдента. Одностороннее **t**-значение может быть получено при замене аргумента вероятность на 2*значимость. Например, для уровня значимости 0,05 и числа степеней свободы 10 двустороннее значение вычисляется СТЮДРАСПОБР(0,05;10). Одностороннее значение для той же значимости и того же числа степеней свободы может быть вычислено формулой СТЮДРАСПОБР(0,05*2;10).

Анализ результатов:

а) Проведем анализ полученных результатов для первой конкурирующей гипотезы $H_1: M_1(X) \neq A = 3,5$ л/мин·м².

Сравним t наблюдаемое и t критическое двустороннее. Так как t наблюдаемое отрицательное, то и t критическое должно быть отрицательным, т. е. равно $-1,98$. Поскольку t наблюдаемое **меньше** t критического ($-8,42 < -1,98$), то **нулевая гипотеза $H_0: M_1(X) = A = 3,5$ л/мин·м² отвергается**, и принимается альтернативная гипотеза $H_1: M_1(X) \neq A = 3,5$ л/мин·м². Это означает, что среднее значение генеральной совокупности критических больных **значимо отличается** от этого же показателя для популяции здоровых людей.

Текст результата анализа: « t набл. $< t$ крит., гипотеза $M_1(X) = 3,5$ отвергается» введите в ячейку (D8).

б) Проведите анализ полученных результатов для **второй** конкурирующей гипотезы $H_2: M_2(X) < A = 3,5$ л/мин·м².

Сравните t наблюдаемое и t критическое **одностороннее**. Поскольку критическая область **левосторонняя**, то t критическое должно быть **отрицательным**, т. е. равно $-1,66$.

Так как t наблюдаемое меньше t критического ($-8,42 < -1,66$), то начальная гипотеза H_0 **отвергается**, и принимается конкурирующая гипотеза H_2 , т. е. среднее значение генеральной совокупности критических больных **значимо меньше** того же показателя для популяции здоровых людей.

Текст результата анализа: « t набл. $< t$ крит., гипотеза $M_1(X) = 3,5$ отвергается и принимается альтернативная гипотеза $M_2(X) < 3,5$ » введите в ячейку (D9).

Задача № 2. По данным предыдущей задачи № 1 рассчитайте доверительный интервал для генерального среднего ($M_1(X)$) у больных в критическом состоянии, при доверительной вероятности $\gamma = 95$ %.

1. Введите в ячейку A10 название вычисляемой характеристики «доверительный интервал».

2. В ячейку B10 вставьте функцию для расчета этой характеристики (рис. 82):

	А	В
10	доверительный интервал	=ДОВЕРИТ(0,05;B5;B3)

Рис. 82

– вставьте функцию **ДОВЕРИТ**, нажав кнопку Вставка функции .

– в окне диалога **Аргументы функции** (рис. 83) в поле *Альфа* введите с клавиатуры уровень значимости **0,05**;

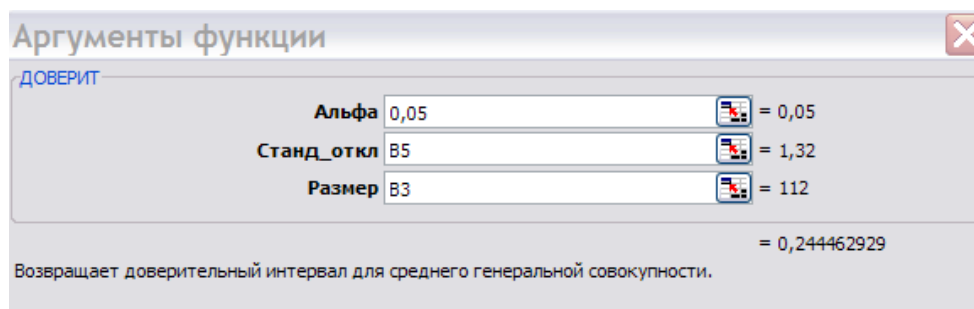


Рис. 83

- в поле *Станд_откл* введите адрес ячейки **B5**, содержащей значение стандартного отклонения;
- в поле *Размер* введите адрес ячейки **B3**, содержащей значение объема выборки; нажмите **ОК**.

	А	В
1	Сердечный индекс	
2	предполагаемое	3,5
3	объём выборки	112
4	среднее выборочное	2,45
5	стандартное отклонение	1,32
6		
7	t наблюдаемое	-8,4183
8	t критическое	1,6587
9		
10	доверительный интервал	0,24446

Рис. 84

Результат всех расчетов задач № 1 и 2 отображен на рис. 84.

3. Введите в ячейку **A11** результат анализа в виде текста «с вероятностью 95 % $M_1(X)$ лежит в интервале $2,45 \pm 0,24$ ».

Таким образом, с вероятностью 95 % интервал $(2,45 \pm 0,24)$ л/мин·м² содержит истинное среднее значение сердечного индекса критически больных пациентов. Предполагаемое $A = 3,5$ л/мин·м², очевидно,

лежит за пределами этого интервала. Поэтому этот интервал сам по себе может использоваться для проверки гипотезы H_0 . Мы отвергаем H_0 с уровнем значимости α , если A лежит вне доверительного интервала для $M_1(X)$.

Задача № 3. Пользуясь разобранным примером (задача № 1), решите следующую задачу: согласно технической норме среднее время выполнения некоторой операции на конвейере 20 с ($A = 20$ с). Поступила жалоба от рабочих, что они затрачивают на эту операцию больше времени. ОТК провел хронометраж у 10 рабочих и получил следующие результаты: $X_b = 22$ с, $S = 3$ с. Можно ли по данным результатам, при уровне значимости 0,02 отклонить исходную гипотезу о том, что среднее время выполнения операции соответствует принятой норме?

Гипотеза H_0 : $M(X) = 20$ с.

Альтернативная гипотеза H_1 : $M(X) \neq 20$ с (двусторонняя критическая область). Уровень значимости $\alpha = 0,02$.

1. Введите данные задачи № 3 согласно образцу на рис. 85.
2. Выполните в ячейке **B19** расчет t -статистики (рис. 86)

	А	В
13	Время выполнения операции	
14	предполагаемое	20
15	объём выборки	10
16	среднее выборочное	22
17	стандартное отклонение	3
18		
19	t наблюдаемое	
20	t критическое	

Рис. 85

	В
19	=(B16-B14)*КОРЕНЬ(B15)/B17
20	=СТЮДРАСПОБР(0,02;B15-1)

Рис. 86

3. Вычислите с помощью статистической функции **СТЮДРАСПОБР** значение t критическое.

4. Результат ваших действий отображен на рис. 87, 88.

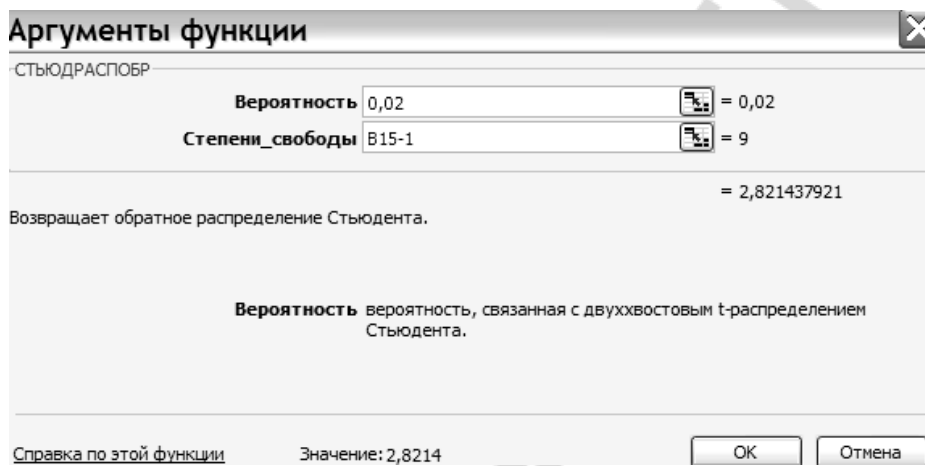


Рис. 87

	А	В
13	Время выполнения операции	
14	предполагаемое	20
15	объём выборки	10
16	среднее выборочное	22
17	стандартное отклонение	3
18		
19	t наблюдаемое	2,11
20	t критическое	2,82

Рис. 88

5. Сделайте вывод о верной статистической гипотезе.

3.7.2. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. Для 8 пациентов с большой массой тела значение общего холестерина (ммоль/л) приведено в таблице. Норма холестерина для людей с нормальным весом составляет 5,5 ммоль/л. Используя данные таблицы, проверьте гипотезу $H_0: M(X) = 5,5$ ммоль/л против конкурирующей гипотезы $H_1: M(X) > 5,5$ ммоль/л, при уровне значимости $\alpha = 0,05$.

Холестерин, ммоль/л	7	8	5,4	6,7	7,6	6,6	8	8,4
---------------------	---	---	-----	-----	-----	-----	---	-----

Задача № 2. Среднее значение гематокрита для здорового взрослого человека равно 40 %. Можно ли по приведенным ниже данным (получены при поступлении в стационар) для пациентов в шоковом состоянии сказать, что у них нормальный гематокрит? Если нет, то определите интервал возможных значений среднего для этой группы больных. Уровень значимости $\alpha = 0,05$.

Гематокрит	28	30	23	48	41	30	27	26	54	27	37	30	27	39
------------	----	----	----	----	----	----	----	----	----	----	----	----	----	----

3.7.3. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ СРЕДНИХ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ, ДИСПЕРСИИ КОТОРЫХ НЕИЗВЕСТНЫ И РАВНЫ, С ИСПОЛЬЗОВАНИЕМ КОМАНДЫ АНАЛИЗ ДАННЫХ

Задача. Исследовалась динамика нарушения ритма по типу желудочковой экстрасистолии у больных острым инфарктом миокарда при их лечении в условиях клиники. Наблюдаемый показатель — количество экстрасистол в час (1/ч). Исходные данные приведены в таблице на рис. 89:

– в контрольной группе — 15 больных, страдающих ишемической болезнью сердца (ИБС);

– в опытной группе — 10 больных с острым инфарктом миокарда на 1, 3 и 9-й день от начала развития заболевания.

	A	B	C	D
1	контроль	1 день	3 день	9 день
2	2	28	15	5
3	5	35	13	3
4	3	40	19	8
5	0	25	5	3
6	1	33	18	7
7	5	42	18	8
8	3	19	5	4
9	2	21	10	5
10	8	28	16	2
11	1	31	15	2
12	0			
13	6			
14	4			
15	2			
16	7			
17				

Рис. 89

Оцените значимость различий показателя в независимых выборках *контроль* и *1-й день (опытная группа)*, *контроль* и *9-й день (опытная группа)*, считая, что в генеральных совокупностях этот показатель распределен по нормальному закону.

Гипотезы H_0 :

1. $M(\text{контр}) = M(1\text{-й день, опыт})$.
2. $M(\text{контр}) = M(9\text{-й день, опыт})$.

Альтернативные гипотезы H_1 :

1. $M(\text{контр}) \neq M(1\text{-й день, опыт})$.
2. $M(\text{контр}) \neq M(9\text{-й день, опыт})$.

Напомним, что если модуль t наблюдаемого $< t$ критического, а $P > \alpha$, то гипотеза H_0 принимается, в противном случае принимается H_1 .

1. Перейдите на новый Лист2. Присвойте листу новое имя «О равенстве средних».

2. Введите данные задачи в таблицу, как показано на рис. 89.
3. Выполните процедуру проверки гипотезы, сравнив **контрольную группу** и **экспериментальную в 1-й день**, для чего:
 - в меню **Сервис** выберите **Анализ данных**;
 - в окне **Анализ данных** (рис. 90) выберите, выделив щелчком мыши, **Двухвыборочный t-тест с различными дисперсиями**; в появившемся окне диалога (рис. 91) в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **A2:A16**, нажмите кнопку разворачивания ;

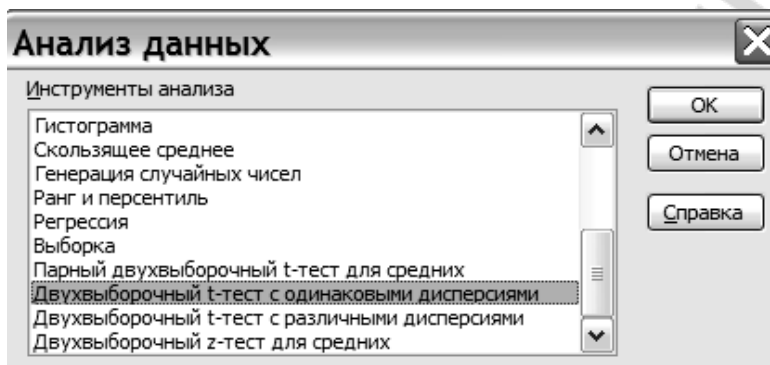


Рис. 90

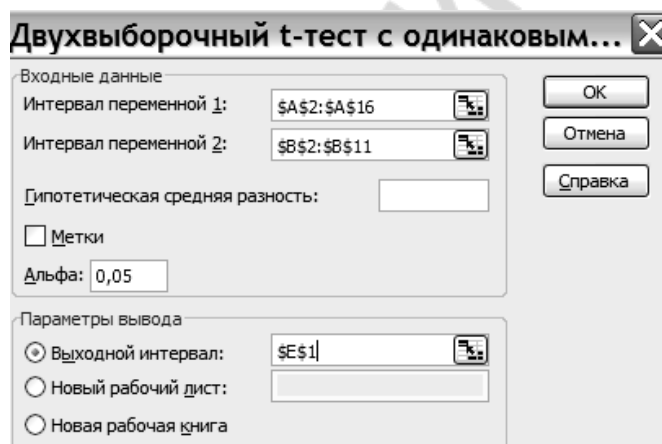


Рис. 91

- в поле *Интервал переменной 2* нажмите кнопку сворачивания  и выделите диапазон ячеек с данными **B2:B11**, нажмите кнопку разворачивания ;
 - в поле *Альфа* введите уровень значимости **0,05**;
 - в поле *Параметры вывода* установите переключатель в позицию *Выходной интервал*;
 - нажмите кнопку сворачивания  и выделите мышью ячейку для вывода результатов **E1**, нажмите кнопку разворачивания ;
 - нажмите **ОК** в окне.
4. В ячейку **E2** введите комментирующий текст «**контроль/1 день**».

5. Выделите заливкой светло-желтого цвета ячейки **F10, F13, F14**, содержащие анализируемые значения.

Результат ваших действий отображен на рис. 92.

	E	F	G
1	Двухвыборочный t-тест с одинаковыми дисперсиями		
2	контроль / 1 день		
3		Переменная 1	Переменная 2
4	Среднее	3,266666667	30,2
5	Дисперсия	6,20952381	57,06666667
6	Наблюдения	15	10
7	Объединенная дисперсия	26,11014493	
8	Гипотетическая разность средних	0	
9	df	23	
10	t-статистика	-12,91103594	
11	P(T<=t) одностороннее	2,5308E-12	
12	t критическое одностороннее	1,713871517	
13	P(T<=t) двухстороннее	5,0616E-12	
14	t критическое двухстороннее	2,068657599	

Рис. 92

Проанализируем результат: t-статистика (наблюдаемое) по модулю больше t критического ($12,9 > 2,069$), а P меньше α ($5,06 \cdot 10^{-12} \ll 0,05$), поэтому гипотеза **H_0** отвергается и принимается гипотеза **H_1** : **$M(\text{контр}) \neq M(1\text{-й день, опыт})$** , **различие** выборочных средних также **значимо**.

6. Результат анализа введите в ячейку **E15**: «t-статистика > t критического, а $P \ll \alpha$, гипотеза о равенстве средних отвергается».

Описанным выше способом выполните проверку гипотезы о равенстве средних для **контрольной группы** и **экспериментальной в 9-й день**.

1. Запустите процедуру проверки, для чего:

- в меню **Сервис** выберите **Анализ данных**;
- в окне **Анализ данных** (см. рис. 90) выберите, выделив щелчком мыши, **Двухвыборочный t-тест с одинаковыми дисперсиями**;
- в появившемся окне диалога в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **A2:A16**, нажмите кнопку разворачивания ;
- в поле *Интервал переменной 2* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **D2:D11**, нажмите кнопку разворачивания ;
- в поле *Альфа* введите уровень значимости **0,05**;
- в поле *Параметры вывода*, установите переключатель в позицию **Выходной интервал**;
- нажмите кнопку сворачивания  и выделите мышью ячейку для вывода результатов **E16**, нажмите кнопку разворачивания ;
- нажмите **ОК** в окне.

2. В ячейку **E17** введите комментирующий текст «контроль/9 день».
3. Выделите заливкой светло-желтого цвета ячейки **F25, F28, F29**, содержащие анализируемые значения.

Результат ваших действий отображен на рис. 93.

	E	F	G
16	Двухвыборочный t-тест с одинаковыми дисперсиями		
17	контроль / 9 день		
18		Переменная 1	Переменная 2
19	Среднее	3,266666667	4,7
20	Дисперсия	6,20952381	5,344444444
21	Наблюдения	15	10
22	Объединенная дисперсия	5,871014493	
23	Гипотетическая разность средних	0	
24	df	23	
25	t-статистика	-1,448992875	
26	P(T<=t) одностороннее	0,080418822	
27	t критическое одностороннее	1,713871517	
28	P(T<=t) двухстороннее	0,160837645	
29	t критическое двухстороннее	2,068657599	

Рис. 93

Проанализируем результат: t-статистика (наблюдаемая) по модулю меньше t критического ($1,45 < 2,07$), а P больше α ($0,16 > 0,05$), поэтому гипотеза о равенстве средних $H_0: M(\text{контр}) = M(9\text{-й день, опыт})$ **принимается**. Выборочные средние статистически не различаются.

4. Введите в ячейку **E31** текст результата анализа: «t-статистика < t критического, а $P > \alpha$, гипотеза H_0 о равенстве средних принимается».

3.7.4. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. По данным задачи, приведенной в 3.7.3, проверьте гипотезу о равенстве средних для двух генеральных совокупностей с различными дисперсиями. Оцените значимость различий исследуемого показателя в независимых выборках «контроль» и «3-й день, опытная группа», считая, что в генеральных совокупностях этот показатель распределен по нормальному закону. Введите в одну из ячеек текст результата анализа, содержащий вывод о равенстве средних.

Задача № 2. Препарат нифедипин обладает способностью расширять сосуды. Его применяют при лечении ИБС. Экспериментально измерялся диаметр коронарных артерий после приема нифедипина и плацебо. В таблице представлены следующие две выборки данных диаметра коронарной артерии (в мм).

Нифедипин	2,5	1,7	1,5	2,5	1,4	1,9	2,3	2,0	2,6	2,3	2,2
Плацебо	2,5	2,2	2,6	2,0	2,1	1,8	2,4	2,3	2,7	2,7	1,9

Позволяют ли приведенные данные полагать, что нифедипин влияет на диаметр коронарных артерий? При решении задачи используйте двухвыборочный t-тест Стьюдента с одинаковыми дисперсиями.

3.7.5. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ ДИСПЕРСИЙ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ С ИСПОЛЬЗОВАНИЕМ КОМАНДЫ АНАЛИЗ ДАННЫХ

В задаче, приведенной в разделе 3.7.3, мы предположили, что дисперсии двух генеральных совокупностей равны. Давайте убедимся в этом с помощью критерия Фишера, используя две независимые выборки «контроль» и «9-й день, опытная группа», при уровне значимости $\alpha = 0,05$.

Гипотеза H_0 : $D(\text{контр}) = D(9\text{-й день, опыт})$. Альтернативная гипотеза H_1 : $D(\text{контр}) > D(9\text{-й день, опыт})$.

Если $F_{\text{наблюдаемое}} < F_{\text{критического}}$ и $P > \alpha$, то нулевая гипотеза о равенстве генеральных дисперсий принимается, выборочные дисперсии также различаются незначимо.

1. Скопируйте в буфер обмена диапазон ячеек **A1:D16**, содержащий выборочные данные.

2. Перейдите на новый рабочий лист **Лист2** и вставьте скопированный диапазон.

3. Переименуйте этот лист, присвоив ему имя **F-тест**.

4. Выполните анализ данных, сравнив контрольную группу и экспериментальную в 9-й день, применив критерий Фишера–Снедекора.

Для этого:

- в меню **Сервис** выберите **Анализ данных**;
- в окне **Анализ данных** (рис. 94) выберите, выделив щелчком мыши, **Двухвыборочный F-тест для дисперсии**;

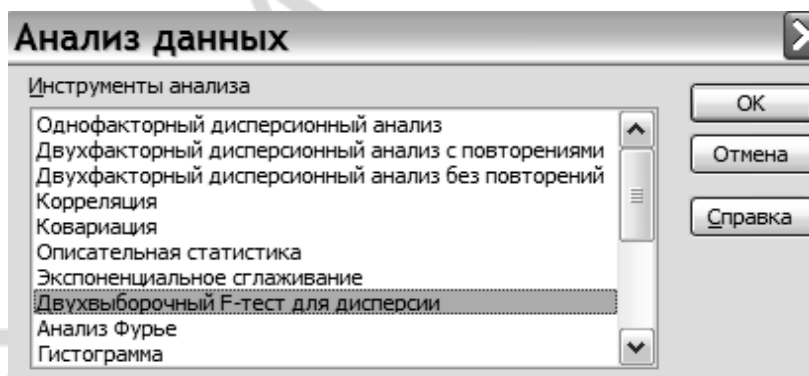


Рис. 94

- в появившемся окне диалога (рис. 95) в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **A2:A16**, нажмите кнопку разворачивания .

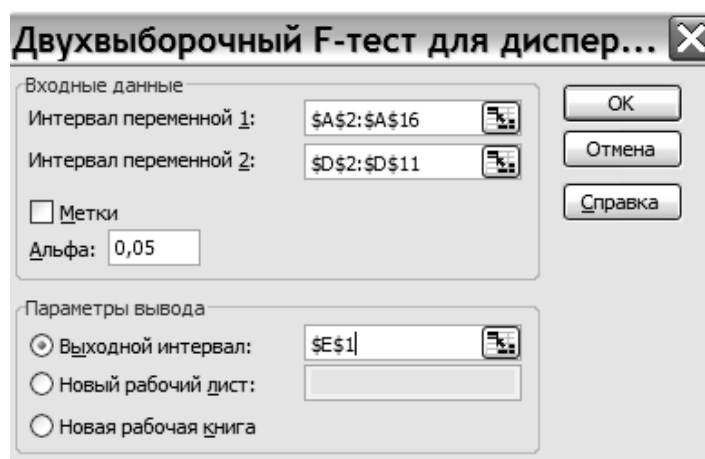


Рис. 95

- таким же образом в поле *Интервал переменной 2* введите диапазон ячеек с данными **D2:D11**;
- в поле *Альфа* введите уровень значимости **0,05**;
- в поле *Параметры вывода*, установите переключатель в позицию *Выходной интервал* и введите в него адрес ячейки **E1** для вывода результатов.
- нажмите **ОК** в окне.

5. В ячейку **E2** введите комментирующий текст «контроль/9 день».

6. Выделите заливкой светло-желтого цвета ячейки **F8, F9, F10**, содержащие анализируемые значения.

Результат ваших действий отображен на рис. 96.

	Е	Ф	Г
1	Двухвыборочный F-тест для дисперсии		
2	контроль/ 9 день		
3		<i>Переменная 1</i>	<i>Переменная 2</i>
4	Среднее	3,266666667	4,7
5	Дисперсия	6,20952381	5,344444444
6	Наблюдения	15	10
7	df	14	9
8	F	1,161865162	
9	P(F<=f) одностороннее	0,42185443	
10	F критическое одностороннее	3,025472724	
11	Т. к. $F_{расч} < F_{кр}$, $P > \alpha$	гипотеза принимается	

Рис. 96

Проанализируем результат: **F**-статистика(наблюдаемое) меньше **F**-критического ($1,16 < 3,03$), а **P** $> \alpha$ ($0,42 > 0,05$), поэтому гипотеза H_0 о равенстве дисперсий **принимается** (дисперсии двух генеральных совокупностей равны, различие генеральных дисперсий не значимо).

7. В ячейку **E11** введите текст результата анализа: «F наблюдаемое < F критического, а $P > \alpha$, гипотеза о равенстве дисперсий принимается».

3.7.6. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. По данным задачи, приведенной в 3.7.3, пользуясь разобранным примером, по двум выборкам (1-й день и 3-й день), при уровне значимости $\alpha = 0,05$, проверьте нулевую гипотезу $H_0: D(1\text{-й день, опыт}) = D(3\text{-й день, опыт})$ о равенстве генеральных дисперсий, при конкурирующей гипотезе $H_1: D(1\text{-й день, опыт}) > D(3\text{-й день, опыт})$. Введите в одну из ячеек текст результата анализа.

Задача № 2. Препарат нифедипин обладает способностью расширять сосуды. Его применяют при лечении ИБС. Экспериментально измерялся диаметр коронарных артерий после приема нифедипина и плацебо. В таблице представлены две выборки данных диаметра коронарной артерии (в мм).

Нифедипин	2,5	1,7	1,5	2,5	1,4	1,9	2,3	2,0	2,6	2,3	2,2
Плацебо	2,5	2,2	2,6	2,0	2,1	1,8	2,4	2,3	2,7	2,7	1,9

Воспользовавшись критерием Фишера, проверьте гипотезу о равенстве дисперсий генеральных совокупностей, соответствующих этим выборкам.

3.7.7. ПРОВЕРКА ГИПОТЕЗЫ О РАВЕНСТВЕ СРЕДНИХ ДВУХ НОРМАЛЬНЫХ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ (ЗАВИСИМЫЕ ВЫБОРКИ) С ИСПОЛЬЗОВАНИЕМ КОМАНДЫ АНАЛИЗ ДАННЫХ

Задача. Воспользуйтесь данными задачи, приведенной в 3.7.3: в опытной группе — 10 больных с острым инфарктом миокарда на 1, 3 и 9-й день от начала развития заболевания. *Исходные данные* — количество экстрасистол в группах (1/ч) (см. рис. 89 и 97).

3-й день	9-й день
15	5
13	3
19	8
5	3
18	7
18	8
5	4
10	5
16	2
15	2

Рис. 97

С помощью парного двухвыборочного t -критерия Стьюдента выполните анализ данных связанных выборок (3-й день, 9-й день) с целью установить равенство средних значений, считая, что в генеральных совокупностях этот показатель распределен по нормальному закону.

Гипотеза $H_0: M(3\text{-й день}) = M(9\text{-й день})$

Альтернативная гипотеза $H_1: M(3\text{-й день}) \neq M(9\text{-й день})$.

Если модуль t наблюдаемого $< t$ критического, а $p > \alpha$, то гипотеза H_0 принимается, в противном случае принимается H_1 .

Для решения задачи выполните следующие действия:

1. Скопируйте с предыдущего листа в буфер обмена диапазон ячеек **C1:D11**, содержащий выборочные данные.

2. Перейдите на новый рабочий лист и вставьте скопированный диапазон (см. рис. 97).
3. Переименуйте этот лист, присвоив ему имя **парный t-тест**.
4. Выполните анализ данных, сравнив **экспериментальные группы на 3-й и 9-й день**, применив парный критерий Стьюдента, для чего:
 - в меню **Сервис** выберите **Анализ данных**;
 - в окне **Анализ данных** (рис. 98) выберите, выделив щелчком мыши, **Парный двухвыборочный t-тест для средних**;

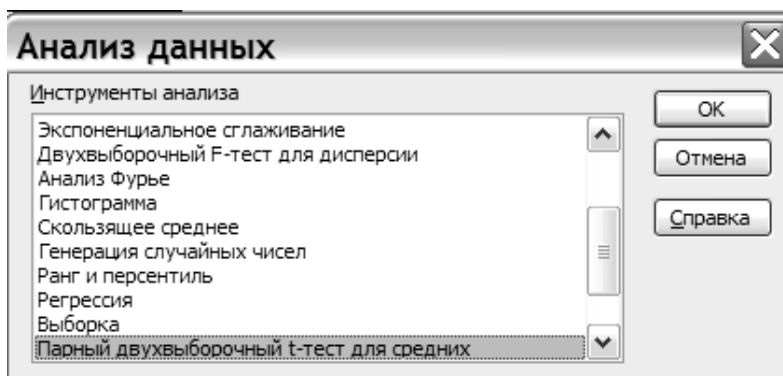


Рис. 98

- в появившемся окне диалога (рис. 99) в группе **Входные данные** поле *Интервал переменной 1* введите диапазон ячеек с данными 3-го дня (**C2:C11**);

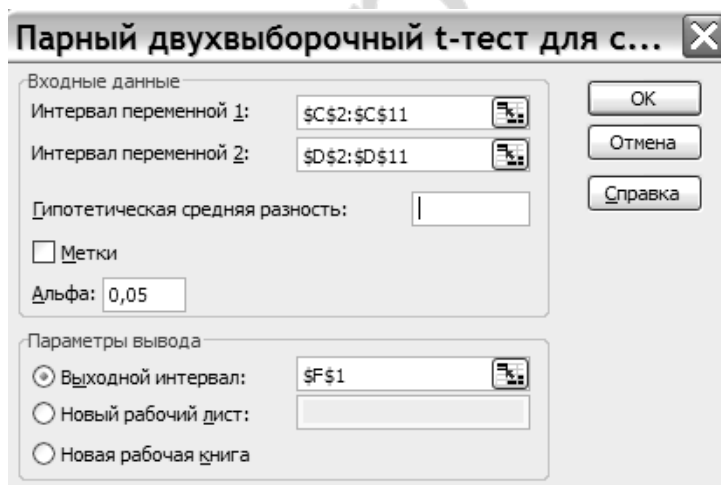


Рис. 99

- в поле *Интервал переменной 2* введите диапазон ячеек с данными 9 дня (**D2:D11**),
- в поле *Альфа* введите уровень значимости **0,05**;
- в поле *Гипотетическая средняя разность* введите **0** или оставьте поле пустым;
- в поле *Параметры вывода*, установите переключатель в позицию *Выходной интервал* и введите в него адрес ячейки вывода результатов **F1**;
- нажмите **ОК** в окне.

5. В ячейку **F2** введите комментирующий текст «3 день/9 день».

6. Выделите заливкой светло-желтого цвета ячейки **G10, G13, G14**, содержащие анализируемые значения.

7. Проанализируем результат: t -статистика (наблюдаемое) больше t критического ($6,15 > 2,26$), а P меньше α ($8,4 \cdot 10^{-5} < 0,05$), поэтому нулевая гипотеза о равенстве средних **отвергается** (средние двух генеральных совокупностей **не равны**).

8. Введите в ячейку **F16** текст результата анализа: « t наблюдаемое $>$ t критического, а $P < \alpha$, гипотеза **H_0** о равенстве средних **отвергается**». Результат ваших действий отображен на рис. 100.

	F	G	H
1	Парный двухвыборочный t-тест для средних		
2	3 день /9 день		
3		Переменная 1	Переменная 2
4	Среднее	13,4	4,7
5	Дисперсия	26,48888889	5,344444444
6	Наблюдения	10	10
7	Корреляция Пирсона	0,496805012	
8	Гипотетическая разность средних	0	
9	df	9	
10	t-статистика	6,150120867	
11	P(T<=t) одностороннее	8,4348E-05	
12	t критическое одностороннее	1,833112923	
13	P(T<=t) двухстороннее	0,000168696	
14	t критическое двухстороннее	2,262157158	
15			
16	t наблюдаемое > t критического, P< α гипотеза отвергается		

Рис. 100

9. Сохраните результаты работы на диске в своей папке.

3.7.8. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. Измерена температура тела у 10 новорожденных детей под мышкой (генеральная совокупность X) и в прямой кишке (генеральная совокупность Y). Проверьте гипотезу о равенстве средних двух этих генеральных совокупностей. Введите в одну из ячеек текст результата анализа.

Задача № 2. Препарат из группы антагонистов кальция — нифедипин — обладает способностью расширять сосуды. Его применяют при лечении ишемической болезни сердца. Экспериментально измерялся диаметр коронарных артерий после приема нифедипина и плацебо. В таблице представлены две выборки данных диаметра коронарной артерии (в мм).

Плацебо	2,5	2,2	2,6	2,0	2,1	1,8	2,4	2,3	2,7	2,7	1,9
Нифедипин	2,5	1,7	1,5	2,5	1,4	1,9	2,3	2,0	2,6	2,3	2,2

Проверьте гипотезу о равенстве средних и ответьте на вопрос: позволяют ли приведенные данные полагать, что нифедипин влияет на диаметр коронарных артерий?

3.8. АНАЛИЗ ДАННЫХ

Основные вопросы:

1. Проверка наличия связи между переменными: расчет коэффициента корреляции.
2. Построение диаграммы рассеяния с помощью Мастера диаграмм.

3.8.1. ПРОВЕРКА НАЛИЧИЯ СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ: РАСЧЕТ ВЫБОРОЧНОГО КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Задача. В 14 анализах крови здоровых пациентов определяли содержание гемоглобина (в %) и оседание эритроцитов крови за 24 часа (в мм).

Гемоглобин, %	22	45	61	66	72	83	73	82	78	82	81	82	77	80
Оседание эритроцитов, мм	8	18	24	26	28	29	30	30	30	32	33	34	35	34

	А	В
1	гемоглобин	оседание эритроцитов
2	22	8
3	45	18
4	61	24
5	66	26
6	72	28
7	83	29
8	73	30
9	82	30
10	78	30
11	82	32
12	81	33
13	82	34
14	77	35
15	80	34

Рис. 101

	А	В	С	Д
1	гемоглобин	оседание эритроцитов		коэффициент корреляции
2				=
3				
4				
5				
6				
7				
8				
9				
10				
11				
12	81	33		
13	82	34		

Рис. 102

Оцените возможную связь между этими параметрами.

Для получения ответа в этом случае воспользуемся Мастером функций.

Выполните следующие действия:

1. Перейдите на новый лист электронной таблицы, присвойте листу имя **Корреляция**.

2. Введите данные задачи в таблицу в виде двух столбцов по образцу на рис. 101. Для этого:

- в ячейку **A1** введите название исследуемого параметра «гемоглобин, %»;
- в ячейку **B1** введите название второго исследуемого параметра «оседание эритроцитов»;
- заполните блок ячеек таблицы **A2:B15** числовыми данными.

3. В ячейку **D1** введите название определяемого параметра «коэффициент корреляции».

– в ячейке **D2** вычислите коэффициент корреляции, используя функцию **КОРРЕЛ** из категории **Статистические**. Подтвердите выбор, нажав **ОК** (рис. 102);

– в появившемся окне *Аргументы функции* (рис. 103) в поле **Массив1** вручную введите адрес первого диапазона ячеек с данными **A2:A15** или выделите этот диапазон в таблице мышью;

– в поле **Массив2** введите адрес второго диапазона ячеек с данными **B2:B15**; подтвердите параметры, нажав **ОК**.

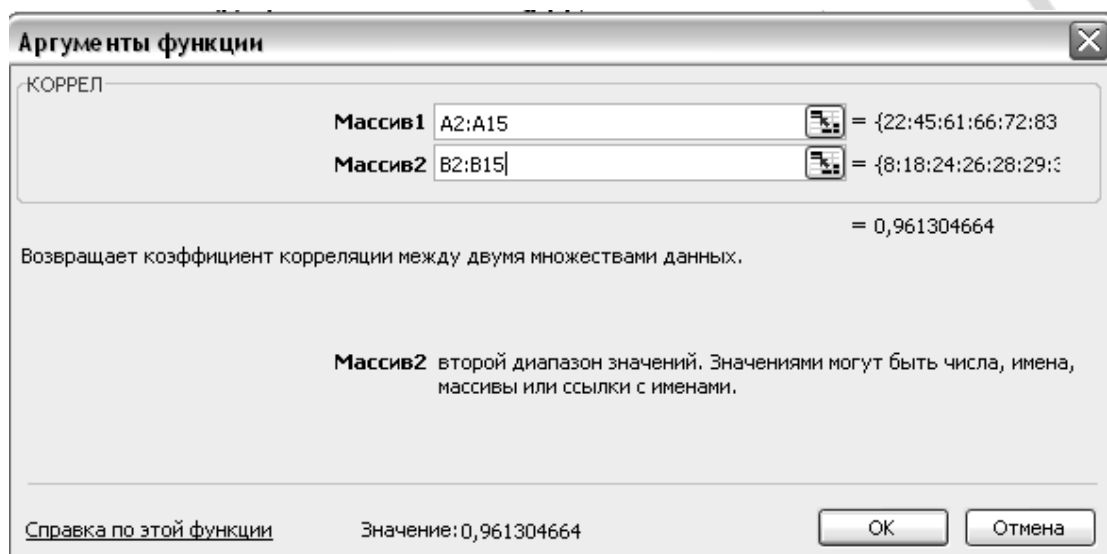


Рис. 103

	A	B	C	D
1	гемоглобин	оседание эритроцитов		коэффициент корреляции
2	22	8		0,961304664
3	45	18		
4	61	24		
5	66	26		
6	72	28		
7	83	29		
8	73	30		
9	82	30		
10	78	30		
11	82	32		
12	81	33		
13	82	34		
14	77	35		
15	80	34		

Рис. 104

Результат ваших действий отображен на рис. 104.

Значение $r = 0,96$, близкое к 1, говорит о наличии тесной линейной связи между параметрами.

3.8.2. ПОСТРОЕНИЕ ДИАГРАММЫ РАССЕЯНИЯ (КОРРЕЛЯЦИОННОГО ПОЛЯ)

Представьте данные задачи, приведенной в подразделе 3.8.1, в виде диаграммы рассеяния (корреляционного поля), иллюстрирующей связь между переменными. Постройте линию регрессии и получите уравнение регрессии.

1. Для построения диаграммы выполните следующие действия:

– выделите диапазон ячеек с данными задачи **A2:B15**; на вкладке **Вставка** в группе **Диаграммы** выберите диаграмму — **Точечная**; затем выберите тип точечной диаграммы → **Точечная с прямыми отрезками и маркерами** (рис. 105);

– используя вкладку **Работа с диаграммами/Макет**, удалите *легенду* и дополните построенную гистограмму следующими элементами: **Название диаграммы** и **Названия осей** как показано на рис. 106.

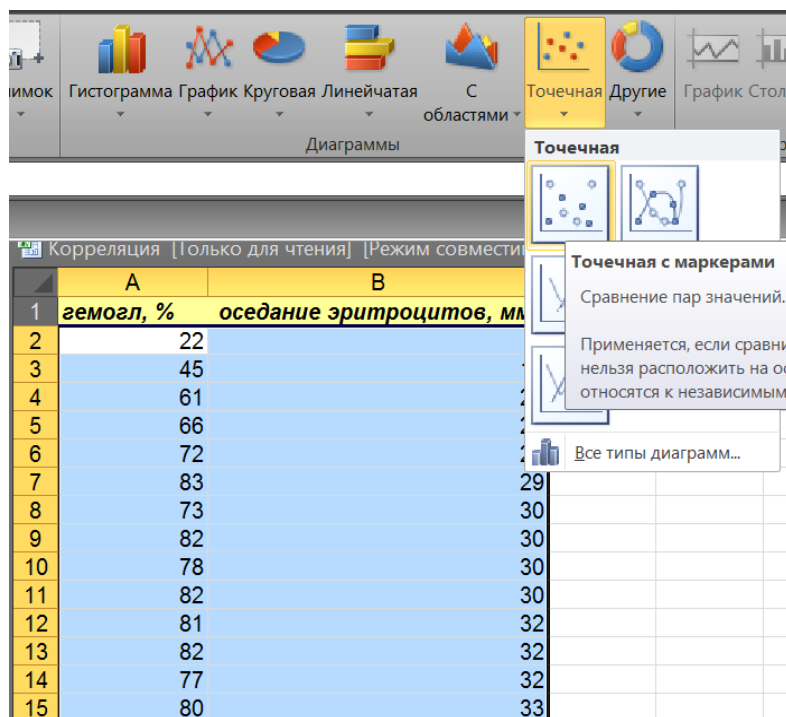


Рис. 105

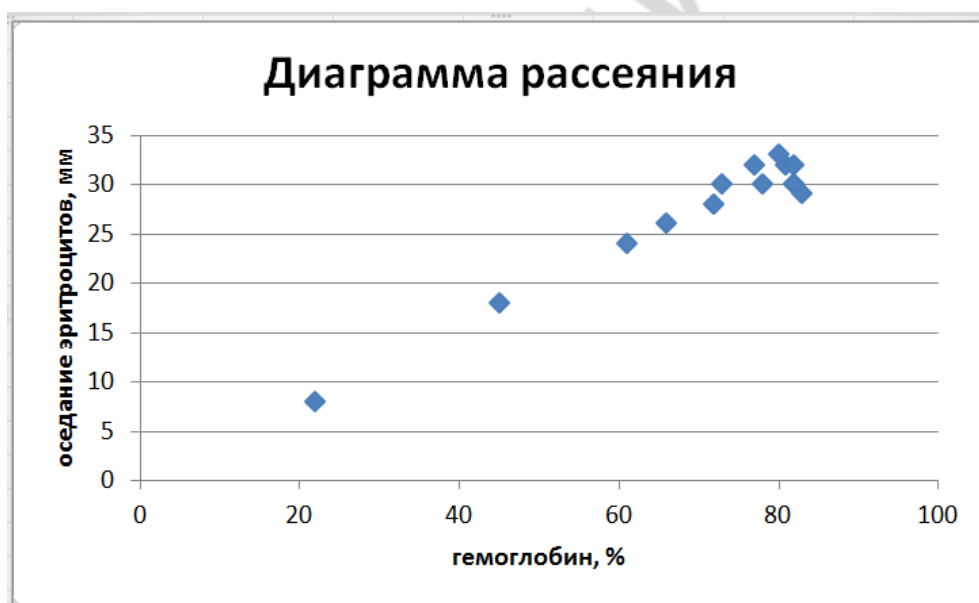


Рис. 106

– Название диаграммы текст **Диаграмма рассеяния**, Ось X (категорий) — текст **гемоглобин, %**; Ось Y (значений) — текст **оседание эритроцитов, мм**;

2. Добавьте линию и уравнение регрессии, для чего:

– наведите указатель мыши на точки диаграммы, выделите их как показано на рис. 107, нажмите правую кнопку мыши для вызова контекстного меню;

– выберите в контекстном меню команду **Добавить линию тренда**;

– в окне диалога **Формат линии тренда** (рис. 108) выберите *Линейная*, поставьте флажки *Показывать уравнение* и *Поместить величину достоверности аппроксимации*¹.

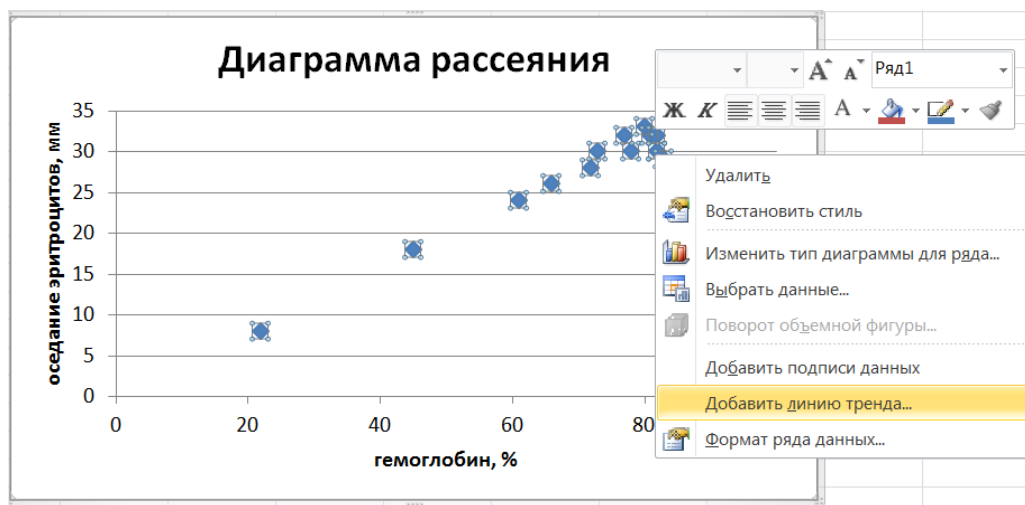


Рис. 107

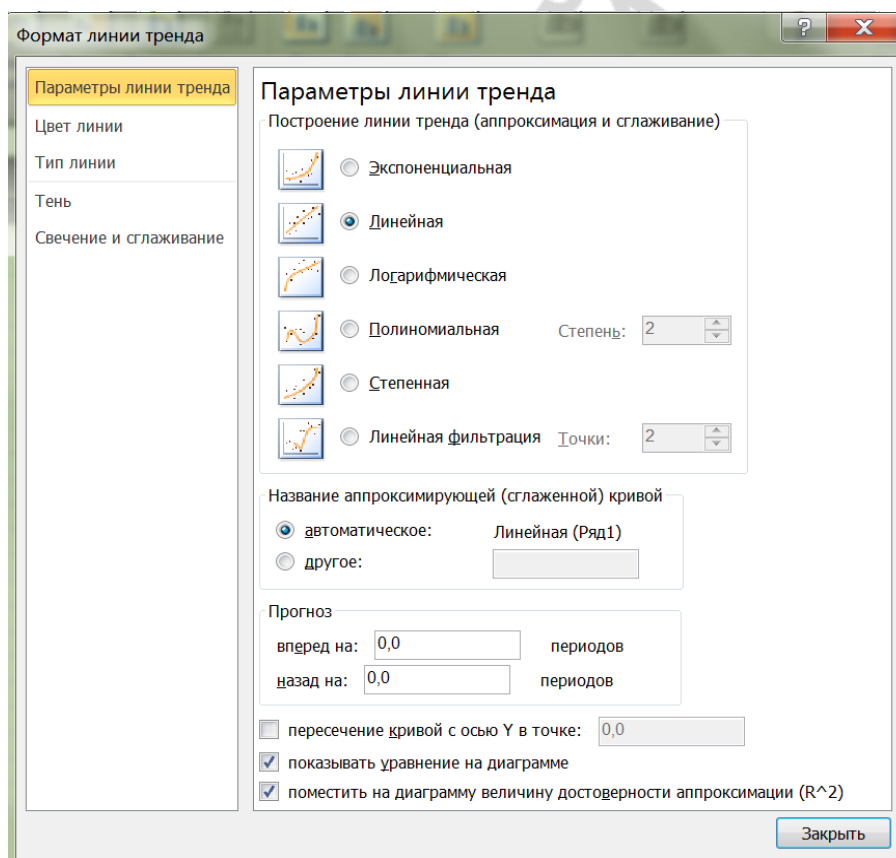


Рис. 108

Результат ваших действий показан на рис. 109.

¹ Достоверность аппроксимации (R -квадрат) отражает близость значений линии тренда к фактическим данным. Линия тренда наиболее соответствует действительности, когда значение R^2 близко к 1.

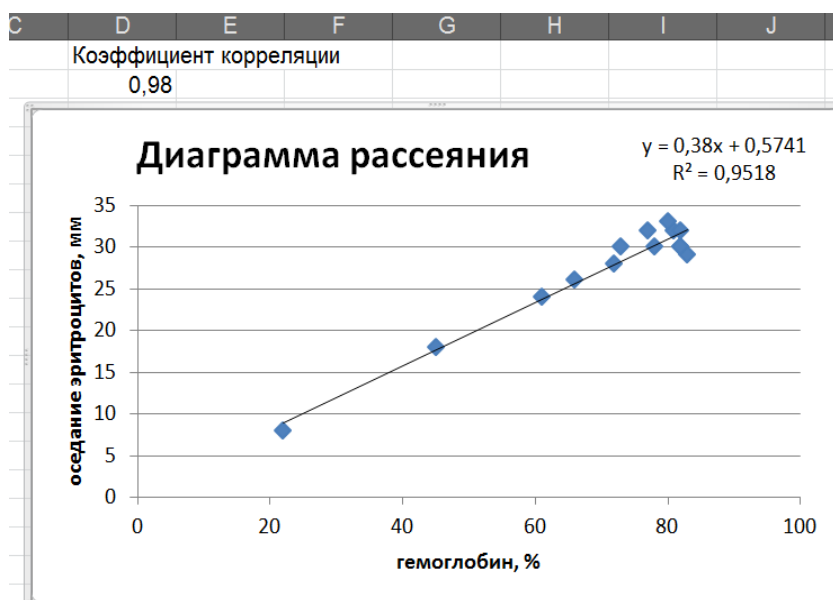


Рис. 109

Так как корреляционное поле вытянуто, то корреляционная связь между признаками есть. Линия регрессии представляет собой прямую, значит корреляционная связь между признаками линейная и оценивается с помощью выборочного коэффициента корреляции r . Так как коэффициент корреляции >0 , то связь между признаками X и Y прямая, т. к. $r = 0,98$ связь сильная.

3.8.3. ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО ВЫПОЛНЕНИЯ

Задача № 1. Для 8 пациентов определена масса тела (кг) и значение общего холестерина (ммоль/л). Используя данные представленные в таблице, оцените возможную связь между этими показателями, построив корреляционное поле и рассчитав выборочный коэффициент корреляции.

Масса, кг	76	87	67	88	85	68	85	115
Холестерин, ммоль/л	7	8	5,4	6,7	7,6	6,6	8	8,4

Задача № 2. В таблице представлены данные по содержанию общего холестерина (ммоль/л) для 11 пациентов страдающих болезнями сердца. Они получены с временным промежутком в 10 лет.

Используя эти данные, постройте корреляционное поле и рассчитайте выборочный коэффициент корреляции. Сделайте вывод о степени связи этих показателей друг с другом.

Холестерин, ммоль/л	Холестерин, ммоль/л (через 10 лет)
7,2	6,5
8,8	7,8
6,3	6,7
7,7	6,7
9,0	6,7
6,9	7,5
8,0	6,0
7,3	6,1
7,8	7,8
9,4	7,1
6,7	7,4

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. *Афифи, А.* Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. Москва : Мир, 1982. 488 с.
2. *Боровиков, В.* Statistica. Искусство анализа данных на компьютере : для профессионалов / В. Боровиков. Санкт-Петербург : Питер, 2001. 656 с.
3. *Вентцель, Е. С.* Теория вероятностей / Е. С. Вентцель. Москва : Наука, 1969. 576 с.
4. *Гмурман, В. Е.* Руководство к решению задач по теории вероятностей и математической статистике / В. Е. Гмурман. Москва : Высшая школа, 2001. 400 с.
5. *Гмурман, В. Е.* Теория вероятностей и математическая статистика / В. Е. Гмурман. Москва : Высшая школа, 1972. 368 с.
6. *Инсарова, Н. И.* Элементы теории вероятностей и математической статистики : учеб.-метод. пособие / Н. И. Инсарова, В. Г. Лещенко. Минск : БГМУ, 2003. 66 с.
7. *Лапач, С. И.* Статистические методы в медико-биологических исследованиях с использованием Excel / С. И. Лапач, А. В. Чубенко, П. И. Бабич. Киев : Морион, 2000. 319 с.
8. *Медик, В. А.* Руководство по статистике здоровья и здравоохранения / В. А. Медик, М. С. Токмачев. Москва : Медицина, 2006. 528 с.
9. *Медик, В. А.* Статистика в медицине и биологии : руководство. В 2 т. Т. 1. Теоретическая статистика / В. А. Медик, М. С. Токмачев, Б. Б. Фишман. Москва : Медицина, 2000. 455 с.
10. *Савіч, Л. К.* Теорыя імавернасцей і матэматычная статыстыка : вучэб. дапаможнік / Л. К. Савіч, К. А. Смольская. Мінск : БДЭУ, 1996. 200 с.
11. *Тюрин, Ю. К.* Статистический анализ данных на компьютере / Ю. Н. Тюрин, А. А. Макаров. Москва : Инфра-М, 1998. 528 с.
12. *Фигурин, В. А.* Теория вероятностей и математическая статистика / В. А. Фигурин, В. В. Оболенкин. Минск : Новое знание, 2000. 206 с.
13. *Юнкеров, В. И.* Математико-статистическая обработка данных медицинских исследований / В. И. Юнкеров, С. Г. Григорьев. Санкт-Петербург : ВМедА, 2002. 266 с.

ОГЛАВЛЕНИЕ

Введение.....	3
Глава 1. Основные понятия теории вероятностей. Случайные величины.	
Законы распределения случайных величин.....	4
1.1. Закономерность и случайность, случайная изменчивость в точных науках, биологии и медицине	4
1.2. Вероятность случайного события	5
1.3. Случайные величины. Виды случайных величин	7
1.4. Закон распределения дискретной случайной величины	7
1.5. Закон распределения непрерывной случайной величины. Плотность распределения вероятностей	8
1.6. Основные числовые характеристики случайных величин	10
1.7. Нормальный закон распределения случайных величин	12
Глава 2. Элементы математической статистики	15
2.1. Предмет и задачи математической статистики. Генеральная и выборочная совокупность	15
2.2. Статистическое распределение выборки.....	16
2.3. Графическое представление статистических распределений выборок.....	19
2.4. Методы описательной статистики	20
2.5. Оценка параметров генеральной совокупности по ее выборке. Точечная и интервальная оценки	23
2.6. Понятие о статистических гипотезах и критериях проверки гипотез.....	26
2.7. Примеры различных критериев и правила работы с ними	30
2.8. Основы корреляционного анализа	37
Глава 3. Анализ данных с помощью табличного процессора Excel	39
3.1. Этапы обработки и анализа экспериментальных данных.....	39
3.2. Описательная статистика. Определение числовых характеристик выборки с помощью формул и Мастера функций.....	39
3.3. Графическое представление статистического распределения выборки....	44
3.4. Применение Пакета анализа для определения числовых характеристик выборки и построения гистограмм.....	57
3.5. Интервальная оценка параметров генеральной совокупности по ее выборке. Расчет доверительных интервалов	62
3.6. Проверка принадлежности распределения выборки к теоретическому нормальному	66
3.7. Проверка гипотез, связанных с параметрами нормального распределения	69
3.8. Анализ данных	86
Список использованной литературы	91