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

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

СТАТИСТИЧЕСКИЙ АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММЫ EXCEL

Учебно-методическое пособие

Минск БГМУ 2010

УДК 577.3 (075.8)
ББК 52.57 я73
Ш46

Рекомендовано Научно-методическим советом университета в качестве учебно-методического пособия 23.06.2010 г., протокол № 11

Р е ц е н з е н т ы: доц. каф. прикладной математики и информатики Белорусского государственного педагогического университета, канд. физ.-мат. наук А. И. Шербаф; доц. каф. медицинской и биологической физики Белорусского государственного медицинского университета, канд. физ.-мат. наук Л. В. Кухаренко

Шеламова, М. А.

Ш46 Статистический анализ медико-биологических данных с использованием программы Excel : учеб.-метод. пособие / М. А. Шеламова, Н. И. Инсарова, В. Г. Лещенко. – Минск : БГМУ, 2010. – 96 с.

ISBN 978-985-528-297-7.

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.

Предназначается для студентов всех факультетов, аспирантов и магистрантов.

УДК 577.3 (075.8)
ББК 52.57 я73

Учебное издание

Шеламова Марина Алексеевна
Инсарова Наталья Ивановна
Лещенко Вячеслав Григорьевич

СТАТИСТИЧЕСКИЙ АНАЛИЗ МЕДИКО-БИОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ ПРОГРАММЫ EXCEL

Учебно-методическое пособие

Ответственный за выпуск В. Г. Лещенко
В авторской редакции
Компьютерная верстка Н. М. Федорцовой
Корректор Ю. В. Киселёва

Подписано в печать 24.06.10. Формат 60×84/16. Бумага писчая «Снегурочка».
Печать офсетная. Гарнитура «Times».
Усл. печ. л. 5,58. Уч.-изд. л. 4,76. Тираж 250 экз. Заказ 734.

Издатель и полиграфическое исполнение:
учреждение образования «Белорусский государственный медицинский университет».
ЛИ № 02330/0494330 от 16.03.2009.
ЛП № 02330/0150484 от 25.02.2009.
Ул. Ленинградская, 6, 220006, Минск.

ISBN 978-985-528-297-7

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, обусловливают 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 … 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:

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

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: $0 \le P(A) \le 1$.

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* — $P^*(A)$:

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

Именно эту величину используют для приближенной оценки статистической вероятности:

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

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

$$P(A) = \lim_{N\to\infty}\left(\frac{M}{N}\right) \qquad (3а)$$

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 нет[1]. 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, pH 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

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

через x_i, а соответствующие им вероятности через p_i[1]. Тогда закон распределения дискретной случайной величины можно задать тремя способами:

1. В виде таблицы, которая называется рядом распределения:

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

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

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

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

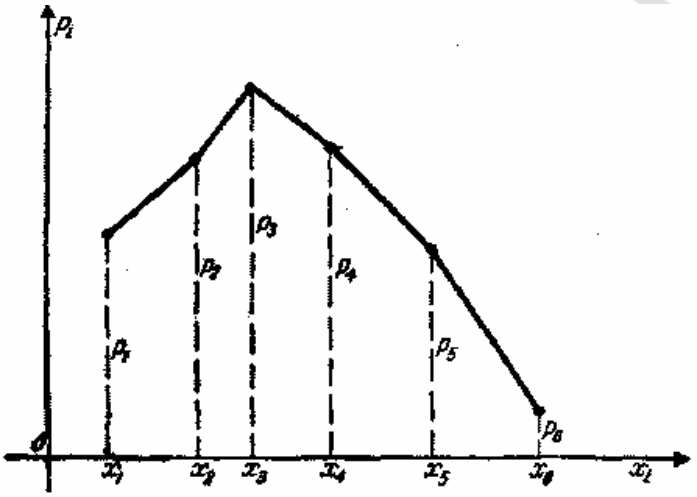


Рис. 1

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-

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

либо ее конкретного значения очень мала (близка к 0)[1]. Вместе с тем 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)[2]. 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).

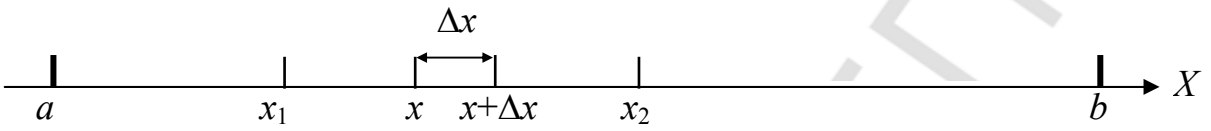


Рис. 2

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

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

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

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

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

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.*

---

[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.

[2] Иногда рассматривают интервал $(-\infty; +\infty)$.

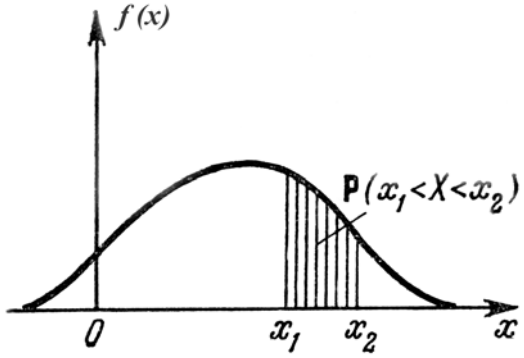


Рис. 3

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

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

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

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

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 $M(X)$.

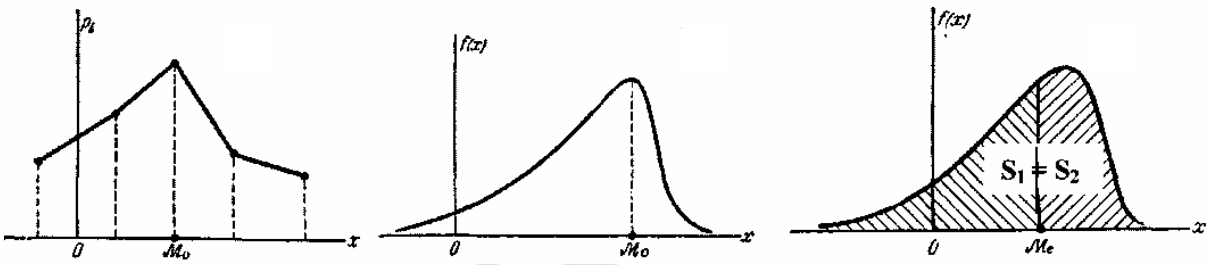
Математическое ожидание $M(X)$ случайной величины X является вероятностным аналогом ее среднего арифметического $\overline{X}$: $M(X) = \overline{X}$ или $M(X) \approx \overline{X}$.

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 $P(X < \text{Me})$ и $P(X > \text{Me})$ 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*).



а *б* *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 ($As < 0$).

Эксцесс 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).

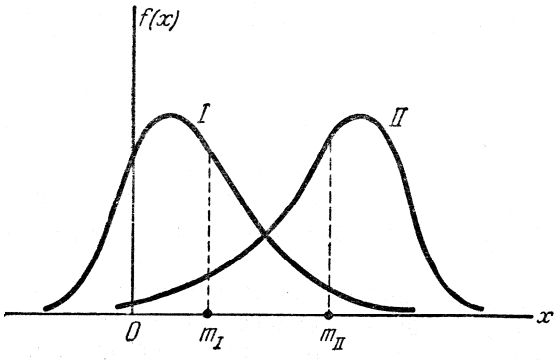


Рис. 5

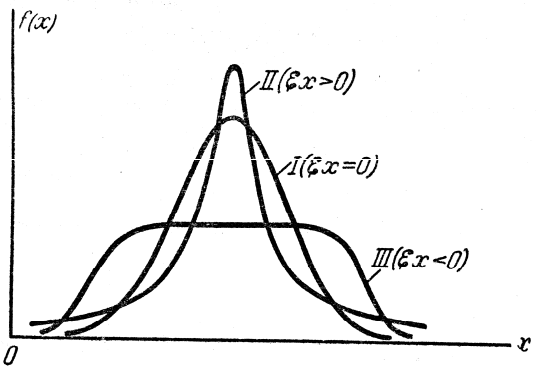


Рис. 6

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:

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

где x — текущие значения случайной величины X; $M(X)$ и σ — ее математическое ожидание и стандартное отклонение. Из (9) видно, что если случайная величина распределена по нормальному закону, то достаточно знать только два числовых параметра — $M(X)$ и σ — чтобы полностью знать закон ее распределения**.**

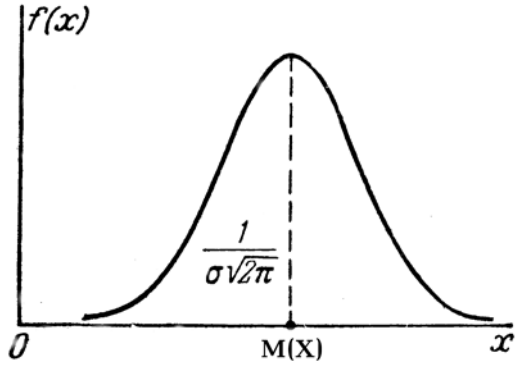


Рис. 7

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 максимальное значение $f(x)$ 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.

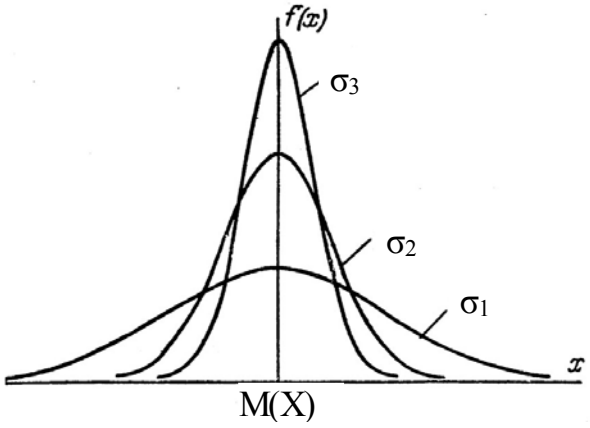


Рис. 8

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:

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

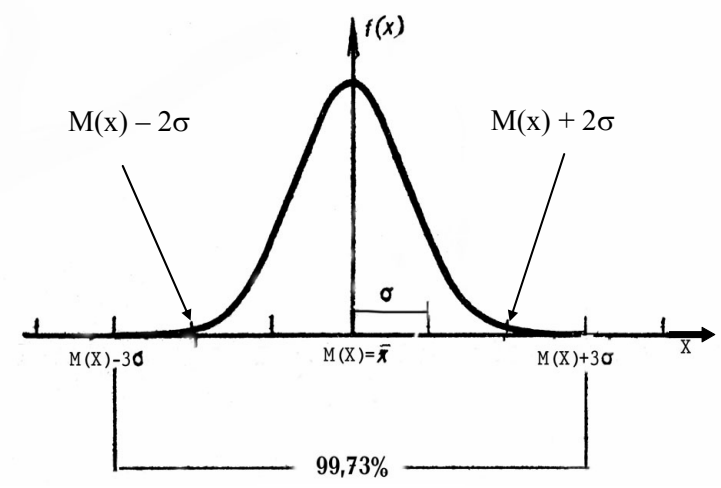


Рис. 9

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:

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

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

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

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 $M(X) = \overline{X}$ 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 pH 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 pH 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 характери-

стике выборки как можно точнее определить соответствующую характеристику 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:* $x_1, x_2, x_3 \ldots x_n$. 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 распределе-

ния 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, … $x_к$ — $m_к$ 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, ранжированной[1], 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 ($x_1 < x_2 < x_3 \ldots < x_k$)	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 ($x_{макс} - x_{мин}$) разбивают на несколько интер-

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

валов обычно равной ширины. Число интервалов k, как правило, не менее 5 и не более 25 и приближенно определяется следующими эмпирическими формулами:

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

где n — объем выборки.

Если ширина интервалов одинакова, то она равна:

$$\Delta x = h = \frac{x_{\text{макс}} - x_{\text{мин}}}{k} \qquad (15)$$

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

Затем для каждого интервала подсчитывают частоты m_i и (или) относительные частоты $p_i = m_i/n$ попадания вариант 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 вариационным 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 статистическим 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); $(x_2, m_2)\ldots (x_к, m_к)$.

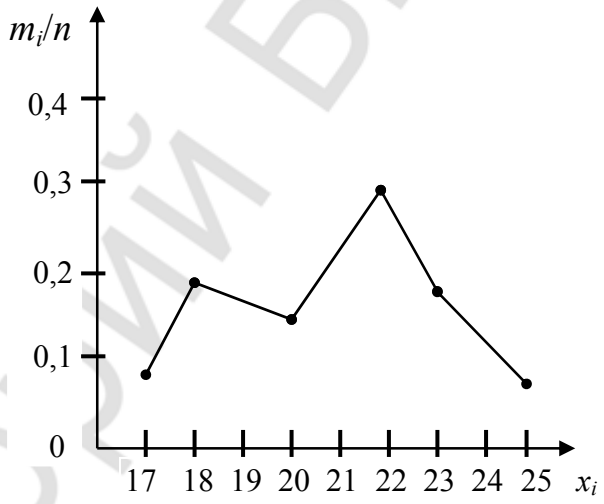


Рис. 10

Полигоном относительных частот называют ломаную линию, отрезки которой соединяют точки $(x_1, \frac{m_1}{n})$; $(x_2, \frac{m_2}{n})$; $(x_к, \frac{m_k}{n})$. На рис. 10 показан полигон относительных частот, построенный по данным выборки для некоторой случайной величины.

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.

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

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

Если строят прямоугольники высотой $\frac{m_i}{\Delta xn}$, 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).

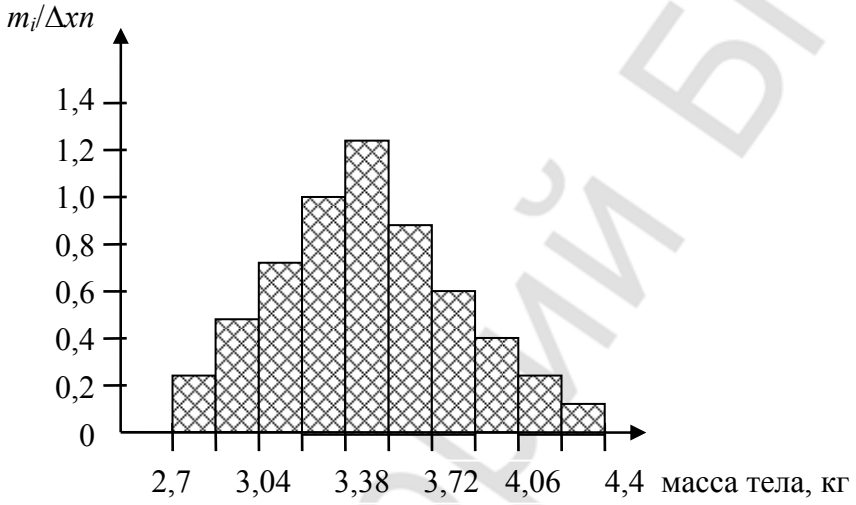


Рис. 11

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 вариант, 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 варианту;

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

$$\bar{x}_{\text{в}} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad (16)$$

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 $Mo_в$ — варианта, которая чаще всего встречается в исследуемой выборке, т. е. имеет наибольшую частоту.

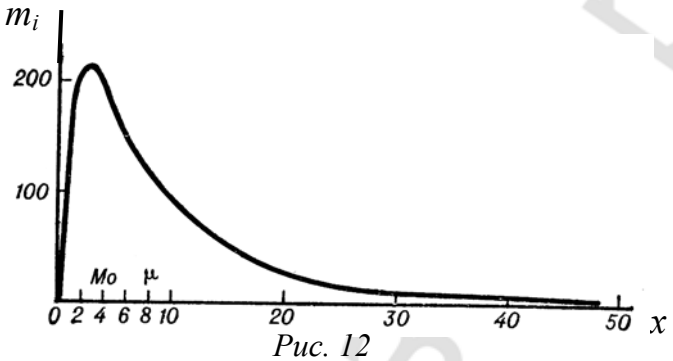


Рис. 12

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 ($\bar{x}_{\text{в}} \approx 7{,}8$ года) в этом случае менее важно, чем знание возраста, в котором чаще всего возникает заболевание и который представляет собой моду ($Mo_{\text{в}} \approx 4$ 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 $Me_в$ — 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. $\text{Me}_\text{в} = 4$.

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

а) *стандартное отклонение* S и *выборочная дисперсия* $D_в = S^2$ [1], характеризующие рассеяние вариант вокруг их среднего выборочного значения $\bar{x}_\text{в}$:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_\text{в})^2}{n-1}}. \qquad (17)$$

Именно эти параметры часто используются как меры изменчивости некоторого исследуемого показателя (случайной величины X). Чем больше $D_\text{в}$ и S, тем сильнее разбросаны значения X относительно среднего;

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

в) *коэффициент вариации*:

$$\nu = \frac{S}{\bar{x}_\text{в}} \cdot 100\ \%. \qquad (18)$$

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

Отметим так же выборочный эксцесс ($Ex_\text{в}$) и выборочную асимметрию $As_\text{в}$ (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 $\bar{x}_\text{в}$, $Me_\text{в}$, $Mo_\text{в}$ равны или незначительно отличаются друг от друга.

2. Минимальное и максимальное значения X ($x_\text{макс}$ и $x_\text{мин}$) примерно равноудалены от $\bar{x}_\text{в}$.

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

3. Выборочные $Ex_в$ и $As_в$ близки к нулю.

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 $\bar{x}_г = M(X)$, генеральные дисперсию $D_г(X)$, стандартное отклонение $\sigma_г$, генеральную моду $Mo_г$, медиану $Me_г$ 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 вероятности 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 $\bar{x}_г = M(X)$ признака, который в генеральной совокупности распределен по нормальному закону. В этом случае интервальной оценкой (с доверительной вероятностью γ) генеральной средней (математического ожидания, $\bar{x}_г = M(X)$) количественного признака X по выборочной средней $\bar{x}_в$ при неизвестном $\sigma_г$ является доверительный интервал

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

или, в другой форме записи:

$$M(X) = \bar{x}_в \pm \delta, \qquad (20)$$

где $\delta = t_{\gamma,n}(S/\sqrt{n})$ — полуширина доверительного интервала — предельная ошибка выборки, характеризующая точность оценки; n — объем выборки; S — выборочное стандартное отклонение; $S/\sqrt{n} = S\bar{x}_в$ — 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 у́же доверительный интервал.

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

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

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

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

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: $\bar{x}_{в} = 58{,}29$ ед., $S = 5{,}58$ ед., $t_{0,9;31} = 1{,}7$. Тогда $M(X) = \bar{x}_{в} \pm t_{\gamma,n} \cdot (S/\sqrt{n}) = (58{,}29 \pm 1{,}7)$ ед. Если указать результат в виде $M(X) = \bar{x}_{в} \pm S/\sqrt{n}$ ($t_{\gamma,n} = 1$), 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 $\bar{x}_{г} = M(X)$.

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 и σ, 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:

$$\frac{(n-1)S^2}{\chi^2_{n-1,(1-\gamma)/2}} < D_{г} < \frac{(n-1)S^2}{\chi^2_{n-1,(1+\gamma)/2}}; \qquad (22)$$

б) для стандартного отклонения:

$$S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,(1-\gamma)/2}}} < \sigma_{г} < S \cdot \sqrt{\frac{(n-1)}{\chi^2_{n-1,(1+\gamma)/2}}}. \qquad (23)$$

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: $\dfrac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} < D < \dfrac{(n-1)S^2}{\chi^2_{n-1,(1-\alpha/2)}}$

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

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 ошибочности 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):

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

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 К, при которых нулевую гипотезу* (H_0) *отвергают.*

2. *Область принятия гипотезы (область допустимых значений) — совокупность значений критерия К, 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.

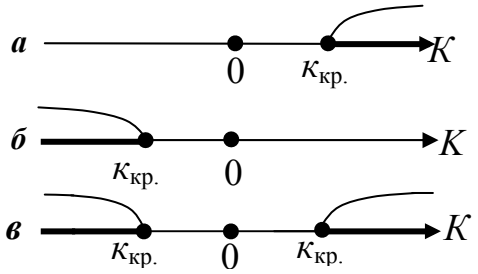


Рис. 13

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 *К*? 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 α = 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 – β называют *мощностью критерия*. Фактически мощность критерия определяется вероятностью не допустить ошибку второго рода ($\beta \to 0$). 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.

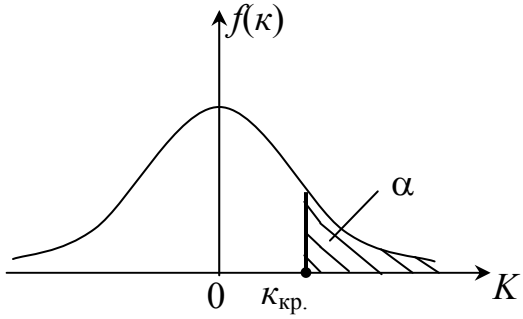


Рис. 14

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

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

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

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

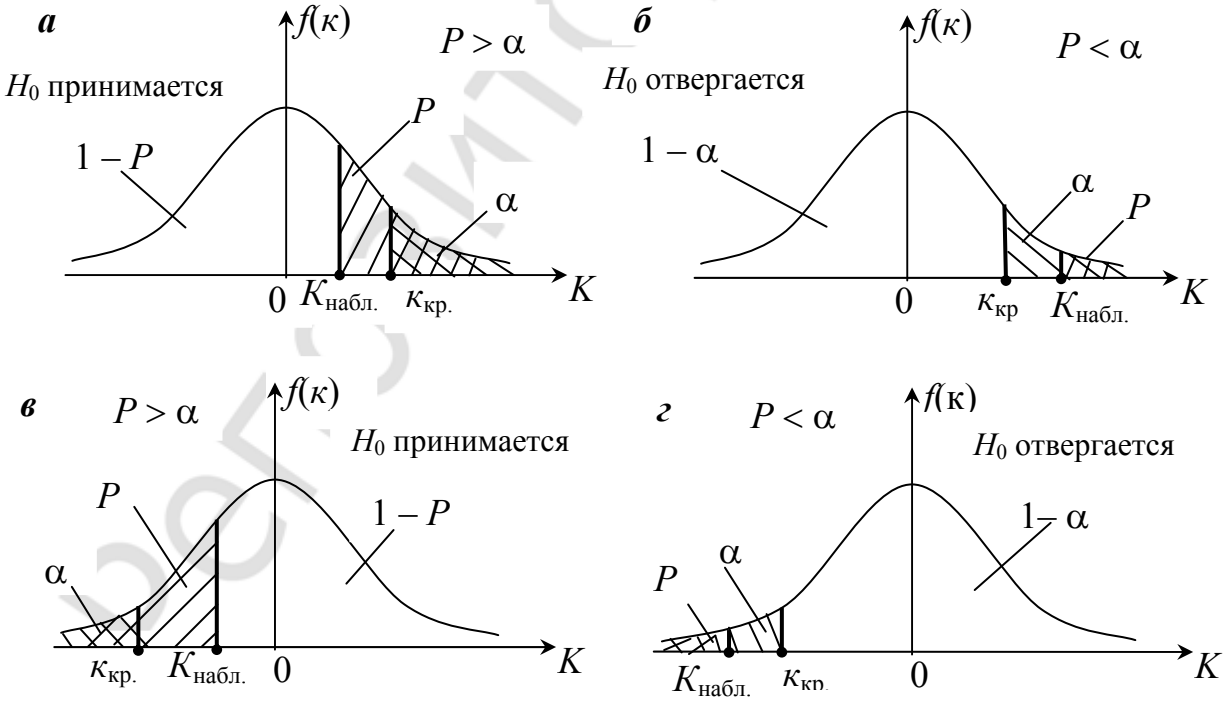


Рис. 15

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

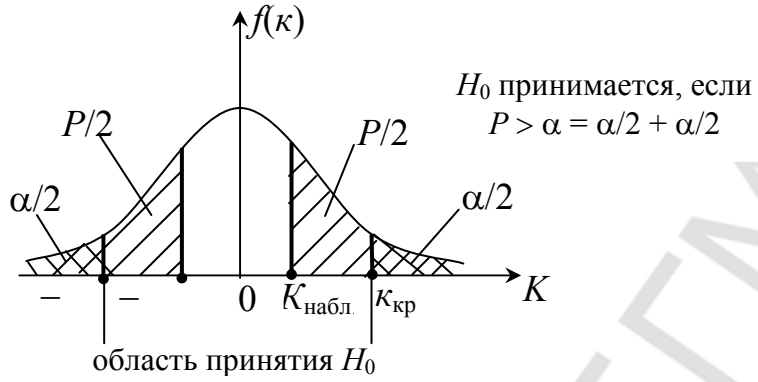


Рис. 16

подход для левосторонней критической области. Этот подход также верен для двусторонней симметричной критической области (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

I. ***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… 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 свободы[1]:

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

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: $M(X) \neq A$ (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 возможность важна

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

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

Если H_0 принимается, т. е. $M(X) = A$, то выборочная средняя $\bar{x}_в$ 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: $\bar{x}_в$ = 35,07 мм, S = 0,16 мм, $t_{набл.}$ = 1,96?

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

Найденное $t_{двухст.\ кр.}(0{,}05;\ 19) = 2{,}09$ и так как $t_{набл.} < t_{двухст.\ кр.}$, то нет оснований при уровне значимости 0,05 отвергать H_0: $M(X)$ = 35 мм, так что отличие $\bar{x}_в$ 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 γ = 1 – α 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 $\bar{x}_в$ = 51,8 см, выборочная дисперсия S^2 = 2,1 см^2, 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 α = 0,05, проверим гипотезу H_0: $M(X)$ = 51,35 см против альтернативы H_1: $M(X) > 51{,}35$ см (правосторонняя критическая область). Вычислим $t_{набл.}$ по (26):

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

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

Так как $t_{набл.} < t_{правост.\ кр.}$, с заданным уровнем значимости α = 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 $\bar{x}_в$ 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 см < $M(X)$ < 52,41 см.

Вывод: этот расчет подтверждает справедливость H_0: с вероятностью 95 % полученный доверительный интервал содержит и средний рост новорожденных А = 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 дисперсиях*[1] *(малые независимые выборки*[2]*, 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 $\bar{x}_в$, $\bar{y}_в$, 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:

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

При $n_1 = n_2 = n$

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

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

[1] Равенство дисперсий можно проверить используя критерий Фишера–Снедекора, который будет рассмотрен 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:

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

$\bar{y}_в = 0{,}03943$ мг/100 г, $S_y = 0{,}00355$ 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 $t_{набл.} = 19{,}6$, $t_{двухст.\ кр.} = 2{,}18$.

Так как $t_{набл.} > t_{двухст.\ кр.}$, 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, при этом 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:

$$F = \frac{S_б^2}{S_м^2} \text{ (критерий } F \text{ Фишера–Снедекора)}, \qquad (30)$$

где $S_м$ и $S_б$ — 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): $F_{\text{набл.}} = \frac{S_б^2}{S_м^2} \geq 1$

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) находят критическую точку $F_{\text{кр.}}(a, m_1, m_2)$. 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 α/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 уро-

вень значимости в два раза меньший заданного, то есть α/2 = 0,025, тогда $F_{кр.} = (0{,}025; 6; 6) = 5{,}82$.

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

3. *Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки[1]).*

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:

$$t = \frac{\overline{d}}{S_d} \cdot \sqrt{n}\,, \tag{31}$$

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

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

Например, если H_1: $M(X) \neq M(Y)$ и $|t_{набл.}| < t_{двухст.\ кр.}$, 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 pH и Y — артериальное pH.

Из клинического опыта известно, что для здоровых людей среднее венозное pH 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: $M(X) < M(Y)$.

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

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

Вывод: в популяции критически больных пациентов среднее венозное pH и среднее артериальное pH значимо отличаются друг от друга, причем $(pH)_{вен.} < (pH)_{арт.}$. 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 корреляционного[1] 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 $x0y$ соответствует одна точка. Всего будет n точек.

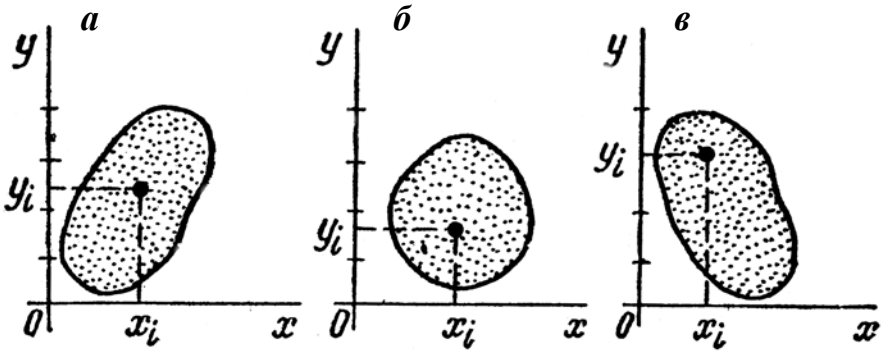


Рис. 17

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

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

По экспериментальным данным, для каждого значения признака X можно найти $\overline{Y}$. Зависимость $\overline{Y}_x = f(x)$ называется *эмпирическим уравнением регрессии Y на X*. Аналогично можно получить зависимость $\overline{X}_y = \varphi(y)$ — *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:

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

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 ($\overline{Y}_x = f(x)$) можно записать в виде:

$$\overline{Y}_x = kx + b, \tag{31}$$

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 $r_г$ нормально распределенной генеральной совокупности (при $n > 50$) можно воспользоваться формулой

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

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 вариации, 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: щелкните левой клавишей мыши на кнопку **Пуск** экрана, выберите в меню **Программы — Microsoft Excel**. В результате на экране появится окно программы, а в нем окно документа **Книга1**.
2. Сохраните созданную книгу под именем **Статистика**, для этого:

– в меню **Файл** выберите команду **Сохранить как**;

– в открывшемся окне сохранения в поле **Папка** укажите имя папки, где должен быть сохранен документ;

– в поле **Имя файла** введите с помощью клавиатуры «**Статистика»**;

– нажмите кнопку **Сохранить**.

3. Переименуйте **Лист1** в **Описательная статистика**:

– подведите курсор к имени листа, щелкните правой кнопкой мыши;

– в контекстном меню выберите **Переименовать**, введите новое имя.

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.

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

– выделите ячейку **A1** щелчком мыши;

– введите в нее название исследуемого параметра «Содержание сывороточного альбумина»;

– завершите ввод, нажав клавишу **Enter**;

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

	A	B	C	D	E	F	G
1	Содержание сывороточного альбумина						Описательная статистика
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		Объём 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							Коэффициент вариации
14							Стандартная ошибка

Рис. 18

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

– с помощью клавиатуры в ячейку **G1** введите заголовок «Описательная статистика», зафиксируйте результат нажатием клавиши **Enter**;

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

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

3.2.2. Расчет выборочных характеристик случайной величины с помощью Мастера функций Excel

Вычисление числовых значений следующих характеристик: выборочного, среднего, медианы, моды, дисперсии, 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 и их обозначением в **Мастере функций** приведено в табл. 3.

Таблица 3

Среднее значение	Мода	Медиана	Минимум	Максимум	Объем выборки	Асимметрия	Эксцесс	Стандартное отклонение	Дисперсия
СРЗНАЧ	МОДА	МЕДИАНА	МИН	МАКС	СЧЕТ	СКОС	ЭКСЦЕСС	СТАНДОТКЛОН	ДИСП

Расчет значения любой из характеристик требует определенной последовательности действий. Разберем ее на следующих примерах:

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**;

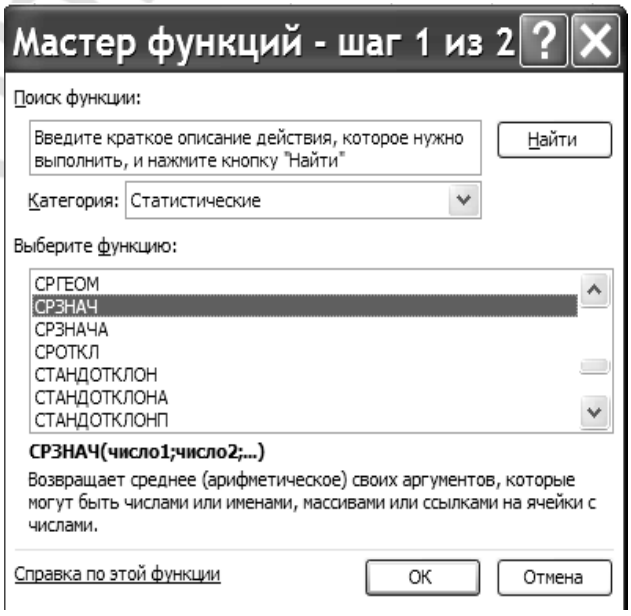


Рис. 19

– 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**;

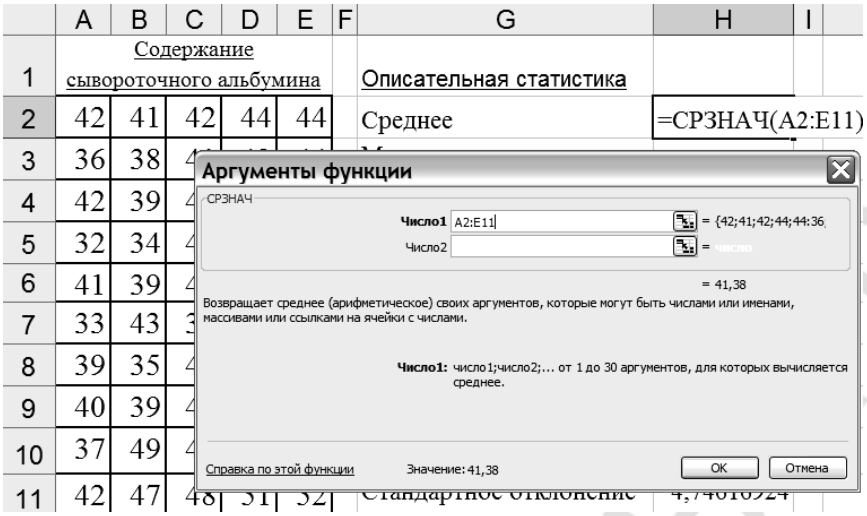


Рис. 20

После этого в ячейке **H2** появится результат расчета среднего значения.

2. Для вычисления моды выполните следующие действия:

– установите курсор в ячейку **H3**;

– нажмите кнопку **Вставка функции**;

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

– в поле *Выберите функцию*, листая список названий функций, найдите и выделите щелчком функцию **МОДА**;

– подтвердите выбор, нажав **ОК**;

– в появившемся окне **Аргументы функции** (рис. 21) в поле *Число1* вручную введите адрес диапазона ячеек с данными **A2:E11** или выделите этот диапазон в таблице мышью;

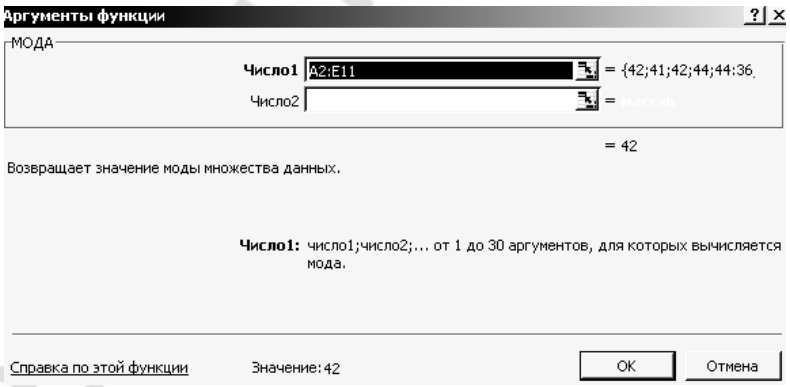


Рис. 21

– подтвердите, нажав **ОК**;

3. Аналогично вычислите в ячейке **H4** значение медианы.

4. Для расчета максимального и минимального значения в ячейки **H5** и **H6** введите функции **МИН** и **МАКС** из категории **Статистические**, соответственно указав в поле *Число1* адрес диапазона ячеек с данными **A2:E11**; подтвердите, нажав **ОК**;

5. Для вычисления объема выборки выполните следующие действия:

– установите курсор в ячейку **H7**;

– нажмите кнопку **Вставка функции**;

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

– в поле ***Выберите функцию***, найдите и выделите щелчком функцию **СЧЕТ**;

– подтвердите выбор, нажав **ОК**;

– в появившемся окне **Аргументы функции** (рис. 22) в поле **Значение1** укажите диапазон ячеек с данными **A2:E11**;

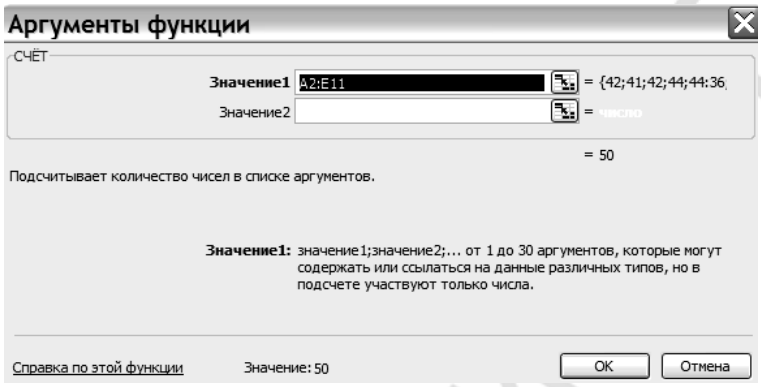


Рис. 22

– подтвердите, нажав **ОК**;

Аналогично рассчитайте значения остальных характеристик: асимметрии и 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, коэффициент вариации, 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 вычисления коэффициента вариации ($\nu = \frac{S}{x_{\text{в}}} \cdot 100\ \%$) — найти отношение стандартного отклонения к среднему значению, для расчета стандартной ошибки ($S/\sqrt{n}$) — разделить стандартное отклонение на корень из объема выборки.

1. Чтобы ***рассчитать вариационный размах*** введите в ячейку **H12** формулу **=H6-H5**, для чего:

– в ячейку **H12** введите с помощью клавиатуры знак = (равно);

– щелкните мышью в ячейке **H6**, где посчитано максимальное значение;

– нажмите на клавиатуре клавишу – (минус);

– щелкните мышью в ячейке **H5**, где посчитано минимальное значение, нажмите клавишу **Enter**.

2. Для ***вычисления коэффициента вариации*** выполните следующие действия:

– установите курсор в ячейку **H13**;

– введите в нее формулу **=H11/H2**, нажмите клавишу **Enter**.

3. Для ***вычисления стандартной ошибки выборочного среднего*** используйте формулу **=H11/КОРЕНЬ(H7)**:

– установите курсор в ячейку **H14** и введите в нее с клавиатуры знак = (равно);

– щелкните мышью в ячейке **H11**, где посчитано стандартное отклонение;

– нажмите на клавиатуре клавишу / (деление);

– для вызова функции КОРЕНЬ нажмите кнопку f_x **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 **H7**; подтвердите, нажав **ОК**.

Окончательный результат ваших действий должен соответствовать рис. 23.

	G	H
1	Описательная статистика	
2	Среднее	=СРЗНАЧ(A2:E11)
3	Мода	=МОДА(A2:E11)
4	Медиана	=МЕДИАНА(A2:E11)
5	Минимум	=МИН(A2:E11)
6	Максимум	=МАКС(A2:E11)
7	Объём выборки	=СЧЁТ(A2:E11)
8	Асимметрия	=СКОС(A2:E11)
9	Эксцесс	=ЭКСЦЕСС(A2:E11)
10	Дисперсия	=ДИСП(A2:E11)
11	Стандартное отклонение	=СТАНДОТКЛОН(A2:E11)
12	Вариационный размах	=H6-H5
13	Коэффициент вариации	=H11/H2
14	Стандартная ошибка	=H11/КОРЕНЬ(H7)

а

	G	H
1	Описательная статистика	
2	Среднее	41,38
3	Мода	42
4	Медиана	42
5	Минимум	32
6	Максимум	52
7	Объём выборки	50
8	Асимметрия	0,02742938
9	Эксцесс	-0,1797134
10	Дисперсия	22,5261224
11	Стандартное отклонение	4,74616924
12	Вариационный размах	20
13	Коэффициент вариации	0,11469718
14	Стандартная ошибка	0,67120969

б

Рис. 23: *а* — в режиме отображения формул; *б* — в режиме отображения результатов расчетов

4. Сохраните результаты работы в том же файле **Статистика.xls** в рабочей папке. Для этого в меню **Файл** выберите команду **Сохранить** или нажмите кнопку **Сохранить**  на Стандартной панели.

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 ± 4,75) г/л; центр рассеяния — среднее 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 $Ex_в$ и $As_в$ 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 x = (МАКС – 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 указываются правые границы интервалов (в Пакете анализа они называются карманы).

3.3.1. Построение гистограммы частот вручную с помощью Мастера функций

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. Найдите в рабочей папке и откройте созданный ранее файл **Статистика.xls** с таблицей «Содержание сывороточного альбумина».

2. Введите в ячейки **B15** и **B16** текст в соответствии с рис. 24, в ячейку **D15** введите заданное по условию число интервалов — 8.

	A	B	C	D
14				
15		число интервалов		8
16		ширина интервала		

Рис. 24

3. В ячейку **D16** вставьте формулу для расчета ширины интервала **=(H6 – H5)/D15** (рис. 25), для чего:

	B	C	D
15	Количество интервалов		8
16	Ширина интервала		=($H$6-$H$5)/D15

Рис. 25

– выделите ячейку **D16**, введите с клавиатуры знак =

(равно) и знак ((скобка);

– щелкните ячейку с максимальным значением **H6**, нажмите клавишу **F4** чтобы сделать ссылку абсолютной;

– введите знак – (минус), щелкните ячейку с минимальным значением **H5**, нажмите клавишу **F4**;

– введите знак) (скобка) и знак / (наклонная черта) и щелкните ячейку **D15** с числом интервалов, зафиксируйте формулу, нажав **Enter**.

	A	B	C	D
14				
15		число интервалов		8
16		ширина интервала		2,5

Рис. 26

Результат ваших действий показан на рис. 26.

4. Составьте таблицу частот исследуемой величины, вводя заголовки столбцов и формулы для расчета границ интервалов и частот (для вычисления правой границы первого интервала используется уравнение **= МИН+x**; для остальных интервалов она определяется путем прибавления величины ширины интервала **x** к правой границе предыдущего):

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

– в блок ячеек **A19:A26** введите номера интервалов (1–8).

5. Вставьте расчетные формулы для вычисления правых границ интервалов в соответствии с рис. 27, для чего:

	A	B	C
18		Интервалы	Частоты
19	1	=$H$5+D16	
20	2	=B19+$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. 27

– установите курсор в ячейку **B19**, нажмите клавишу = (равно);

– укажите мышью на ячейку **H5**, где посчитано минимальное значение, нажмите клавишу **F4**, чтобы сделать ссылку абсолютной, затем нажмите клавишу + (плюс);

– укажите мышью на ячейку **D16**, где посчитана ширина интервала, нажмите **Enter**;

– установите курсор в ячейку **B20**, нажмите клавишу = (равно);

– укажите мышью на ячейку **B19**, где посчитана граница первого интервала, нажмите клавишу + (плюс);

– укажите на ячейку **D16**, где вычислена ширина интервала, нажмите клавишу **F4**, чтобы сделать ссылку абсолютной, нажмите **Enter**.

6. Растяните формулу из ячейки **B20** при помощи автозаполнения, в ячейки от **B21** до **B26** для чего, установив курсор в ячейку **B20**, укажите на правый нижний угол этой ячейки до появления маркера автозаполнения + (плюс), нажмите левую кнопку мыши и, удерживая ее, протяните выделение до ячейки **B26** и отпустите кнопку мыши.

7. Вставьте в блок ячеек **C19:C26** функцию для расчета частот, указывая массив данных и массив интервалов[1]. Для этого:

– установите курсор в ячейку **C19**;

– вызовите **Мастер функций,** нажав кнопку f_x;

– найдите и выберите из категории **Статистические** функцию **ЧАСТОТА**, подтвердите, нажав **ОК**;

– в окне **Аргументы функции** (рис. 28) в поле **Массив_данных**, нажмите кнопку сворачивания и выделите мышью диапазон ячеек исходных данных **A2:E11**, затем нажмите кнопку разворачивания;

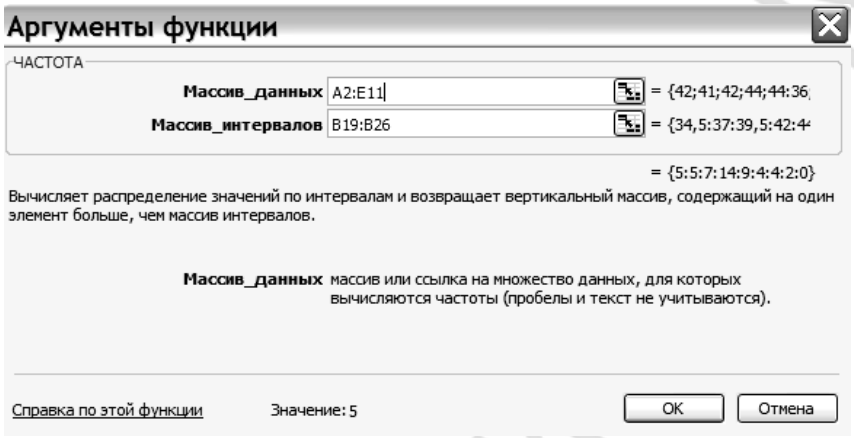


Рис. 28

– справа от поля **Массив_интервалов** нажмите кнопку сворачивания, выделите мышью диапазон ячеек интервалов **B19:B26**, нажмите кнопку разворачивания, подтвердите, нажав **ОК**;

– 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. 29.

	B	C	D
14			
15	число интервалов		8
16	ширина интервала		2,5
17			
18	Интервалы	Частоты	
19	34,5	5	
20	37	5	
21	39,5	7	
22	42	14	
23	44,5	9	
24	47	4	
25	49,5	4	
26	52	2	

Рис. 29

8. Постройте гистограмму для исследуемой величины с применением Мастера диаграмм, для чего:

– выделите диапазон ячеек с таблицей частот **C19:C26**;

– щелкните мышью кнопку **Мастер диаграмм**;

– на вкладке **Стандартные** в поле *Тип* выберите вариант *Гистограмма* и нажмите кнопку **Далее**;

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

– в появившемся окне на вкладке **Диапазон данных** включите переключатель *Ряды в: столбцах* 

;

– на вкладке **Ряд** (рис. 30) в поле **Подписи оси X** нажмите кнопку сворачивания  , выделите диапазон ячеек **B19:B26**, нажмите кнопку разворачивания  , нажмите кнопку **Далее>**;

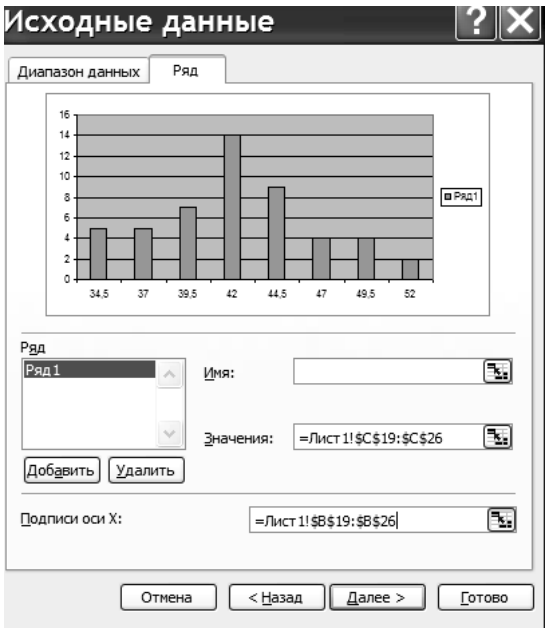


Рис. 30

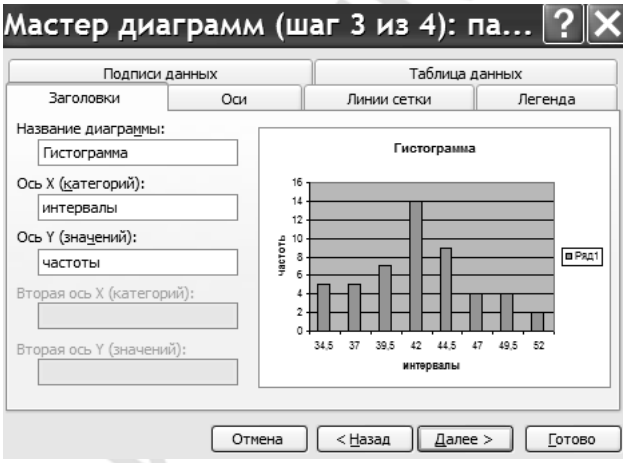


Рис. 31

– в следующем окне на вкладке **Заголовки** (рис. 31) напечатайте в полях *Название диаграммы* текст **Гистограмма**, *Ось X (категорий)* — текст **интервалы**; *Ось Y (значений)* — текст **частоты**;

– на вкладке **Линии сетки** (рис. 32) установите флажки *основные линии* в разделах **Ось X** и **Ось Y**;

– на вкладке **Легенда** снимите флажок *Добавить легенду*, нажмите кнопку **Далее>**;

– в появившемся окне выбора места расположения диаграммы (рис. 33) включите переключатель **имеющемся** и нажмите кнопку **Готово**.

Рис. 32

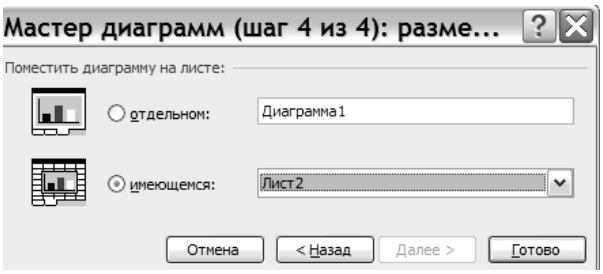


Рис. 33

Результат ваших действий показан на рис. 34.

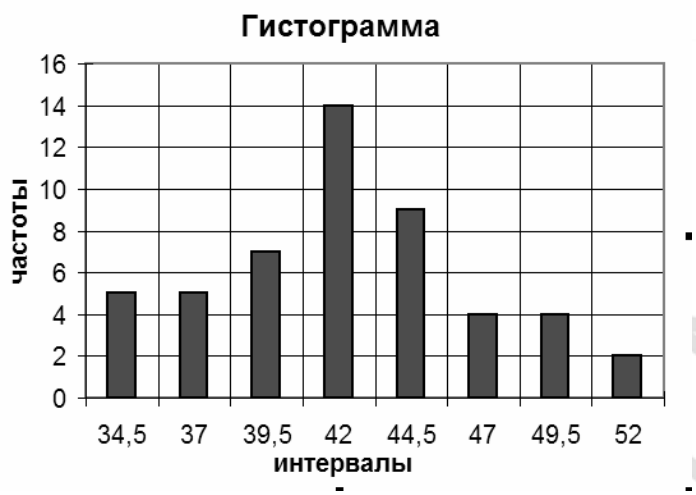


Рис. 34

9. Сохраните повторно в рабочую папку файл **Статистика.xls** с результатами вашей работы.

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. Число интервалов $n = 8$.

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

Для решения задачи выполните следующие действия:

1. Дополните таблицу частот исследуемой величины по образцу на рис. 35, выполнив следующие действия:

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

– введите в ячейку **D19** формулу для расчета плотности относительной частоты **=C19/H10/D16**;

– нажмите клавишу **F4**, чтобы сделать ссылки **H10** и **D16** абсолютными.

2. Скопируйте формулу из ячейки **D19** при помощи автозаполнения, в диапазон ячеек **D20:D26**, для чего укажите на правый нижний угол ячейки **D19** до появления маркера автозаполнения + (плюс), нажмите левую кнопку мыши и, удерживая ее, протяните выделение до ячейки **D26** и опустите кнопку мыши.

Результат ваших действий отображен на рис. 36.

	D
18	Плотность относ частоты
19	=C19/$H$10/$D$16
20	=C20/$H$10/$D$16
21	=C21/$H$10/$D$16
22	=C22/$H$10/$D$16
23	=C23/$H$10/$D$16
24	=C24/$H$10/$D$16
25	=C25/$H$10/$D$16
26	=C26/$H$10/$D$16

Рис. 35

	A	B	C	D
18		Интервалы	Частоты	Плотность относ частоты
19	1	34,5	5	0,04
20	2	37	5	0,04
21	3	39,5	7	0,056
22	4	42	14	0,112
23	5	44,5	9	0,072
24	6	47	4	0,032
25	7	49,5	4	0,032
26	8	52	2	0,016
27			50	0,4

Рис. 36

3. Можно выполнить проверку расчетов, вычислив с помощью кнопки Автосумма Σ на Стандартной панели, в ячейках **C27** и **D27** суммы частот и плотностей относительных частот: $\sum_{i=1}^{k} m_i = n$ и $\sum_{i=1}^{k} \frac{m_i}{n \cdot \Delta x} = \frac{1}{\Delta x}$.

4. Постройте гистограмму относительных частот с применением Мастера диаграмм для чего:

– выделите диапазон ячеек с таблицей плотности относительных частот **D19:D26**;

– щелкните мышью кнопку **Мастер диаграмм**;

– на вкладке **Стандартные** в поле *Тип* выберите вариант *Гистограмма* и нажмите кнопку **Далее**;

– в появившемся окне на вкладке **Диапазон данных** включите переключатель *Ряды в: столбцах*;

– на вкладке **Ряд** (рис. 37) в поле **Подписи оси X** нажмите кнопку сворачивания, выделите диапазон ячеек **B19:B26**, нажмите кнопку разворачивания, нажмите кнопку **Далее>**;

– в следующем окне на вкладке **Заголовки** напечатайте в полях *Название диаграммы* текст

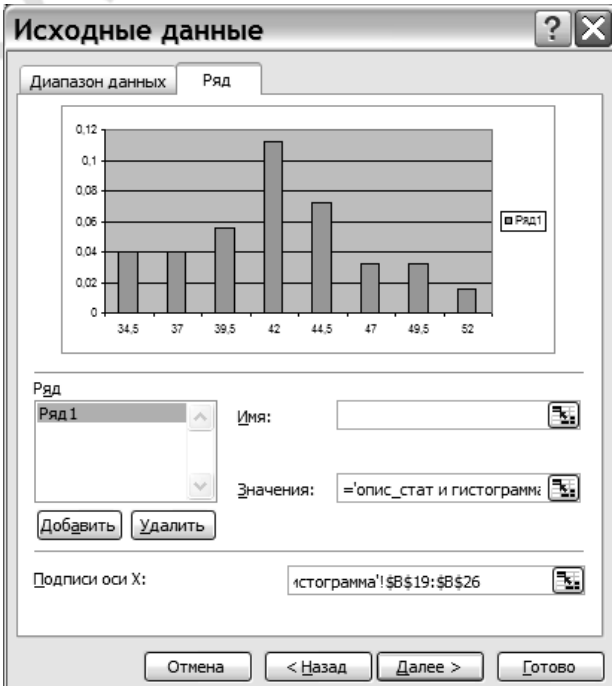


Рис. 37

Гистограмма относительных частот, *Ось X (категорий)* — текст **интервалы**, *Ось Y (значений)* — текст **плотность относительной частоты**;

– на вкладке **Линии сетки** установите флажки *основные линии* в разделах **Ось X** и **Ось Y**;

– на вкладке **Легенда** снимите флажок *Добавить легенду*, нажмите кнопку **Далее>**;

– в появившемся окне выбора места расположения диаграммы включите переключатель **имеющемся** и нажмите кнопку **Готово**.

Результат ваших действий показан на рис. 38.

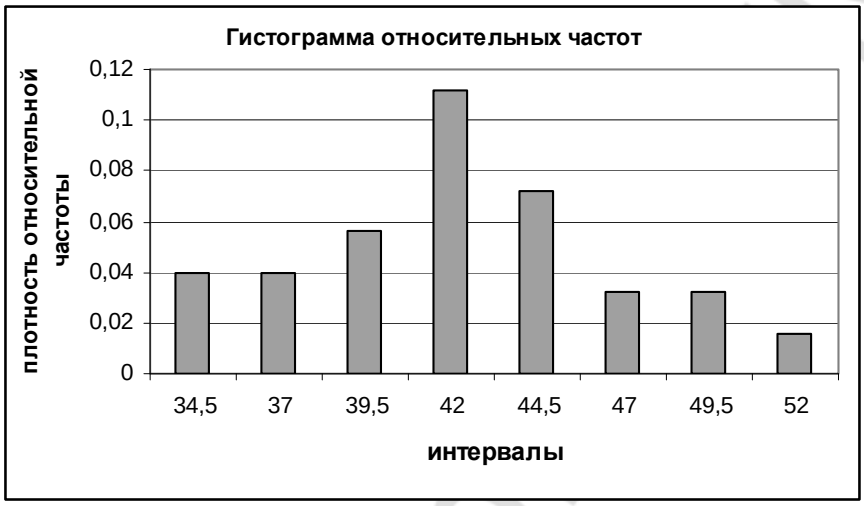


Рис. 38

3.3.3. Форматирование гистограммы

1. Скорректируйте первую построенную гистограмму частот:

– выделите диаграмму щелчком мыши по ней;

– укажите мышью на угловой ограничительный маркер диаграммы до появления указателя в форме двунаправленной стрелки и растяните мышью размеры диаграммы;

– выделите щелчком мыши область построения гистограммы; установите указатель на центральный маркер верхней границы области, переместите рамку диаграммы выше, увеличив ее в высоту.

2. Измените вид прямоугольников гистограммы:

– щелкните правой кнопкой мыши на прямоугольниках гистограммы;

– выберите в появившемся контекстном меню команду **Формат ряда данных**;

– на вкладке **Вид** в области *Заливка* выберите щелчком в палитре другой цвет для прямоугольников;

– на вкладке **Параметры** (рис. 39) в поле *Ширина зазора* установите число «0», нажмите **ОК**.

3. Снимите заливку области построения диаграммы, для чего:

– выделите область построения (сетку) щелчком правой кнопки мыши;

– в контекстном меню выберите **Формат области построения**;

– на вкладке **Вид** (рис. 40) в области *Заливка* над палитрой установите переключатель прозрачная.

– в области *Рамка* установите переключатель невидимая, нажмите **ОК**.

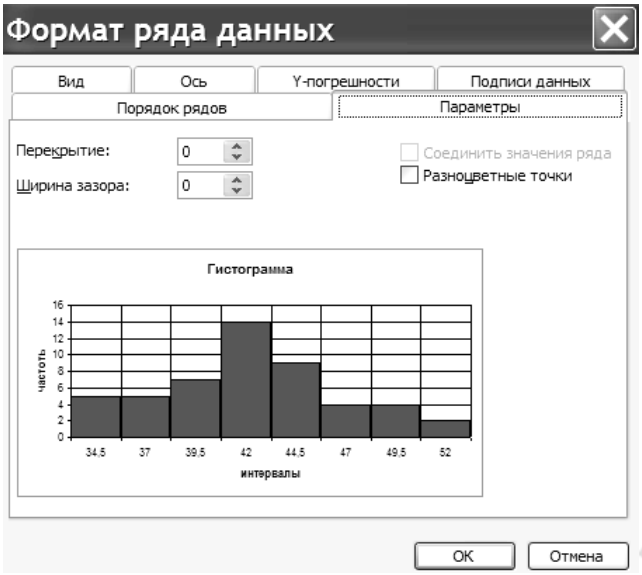


Рис. 39

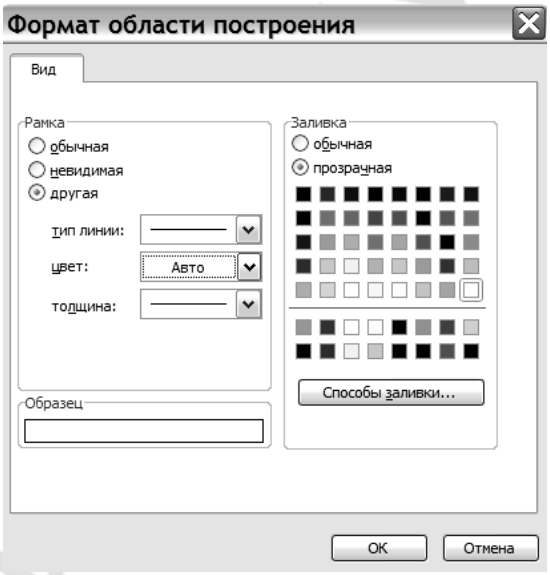


Рис. 40

Результат ваших действий представлен на рис. 41.

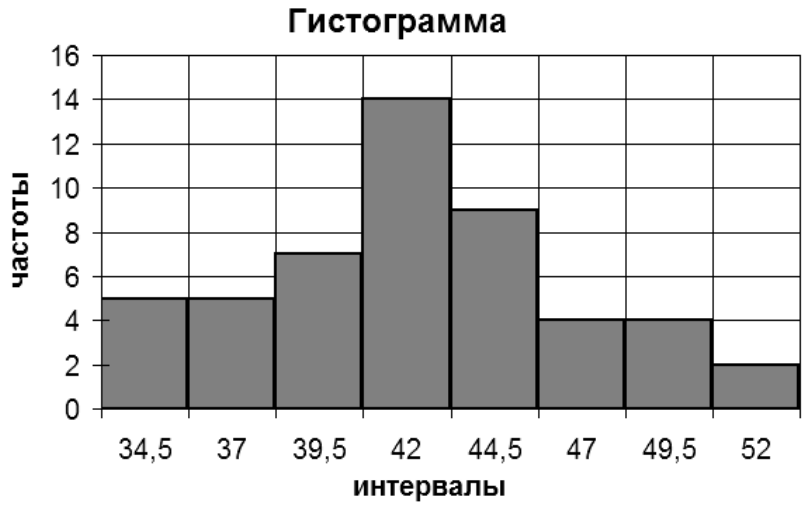


Рис. 41

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**;
– выберите в меню **Правка** команду **Копировать**;
– щелкните в ячейке **A33**;
– выберите в меню **Правка** команду **Вставить**.

2. Скорректируйте скопированный фрагмент (рис. 42): в ячейке **D33** удалите старое содержимое клавишей **Del** и введите число **5**.

	A	B	C	D
33		Количество интервалов		5
34		Ширина интервала		=($H$6-$H$5)/D33
35				
36		Интервалы	Частоты	
37	1	=H5+D34	=ЧАСТОТА(A2:E11;B37:B41)	
38	2	=B37+$D$34	=ЧАСТОТА(A2:E11;B37:B41)	
39	3	=B38+$D$16	=ЧАСТОТА(A2:E11;B37:B41)	
40	4	=B39+$D$16	=ЧАСТОТА(A2:E11;B37:B41)	
41	5	=B40+$D$16	=ЧАСТОТА(A2:E11;B37:B41)	

Рис. 42

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

4. Измените формулы расчета границ интервалов:

– щелкните мышью в ячейке **B37**;
– установите курсор в строке формул;
– удалите ссылку **H23** и введите вместо нее ссылку **H5**, нажмите **Enter**;
– щелкните мышью в ячейке **B38**;
– установите курсор в строке формул;
– исправьте ссылку **D16** на абсолютную **D34**, нажмите **Enter**;
– скопируйте формулу из ячейки **B37** в диапазон **B38:B41**, убедитесь, что произошел перерасчет значений правых границ интервалов.

5. Вставьте в блок ячеек **C37:C41** функцию для расчета частот. Для этого:

– установите курсор в ячейку **C37**;
– вызовите **Мастер функций**, нажав кнопку f_x;
– найдите и выберите из категории **Статистические** функцию **ЧАСТОТА**, подтвердите, нажав **ОК**;
– в открывшемся окне **Аргументы функции** в поле **Массив_данных** нажмите кнопку сворачивания;
– выделите мышью диапазон ячеек **A2:E11**, нажмите кнопку разворачивания;

– в поле **Массив_интервалов** нажмите кнопку сворачивания ;

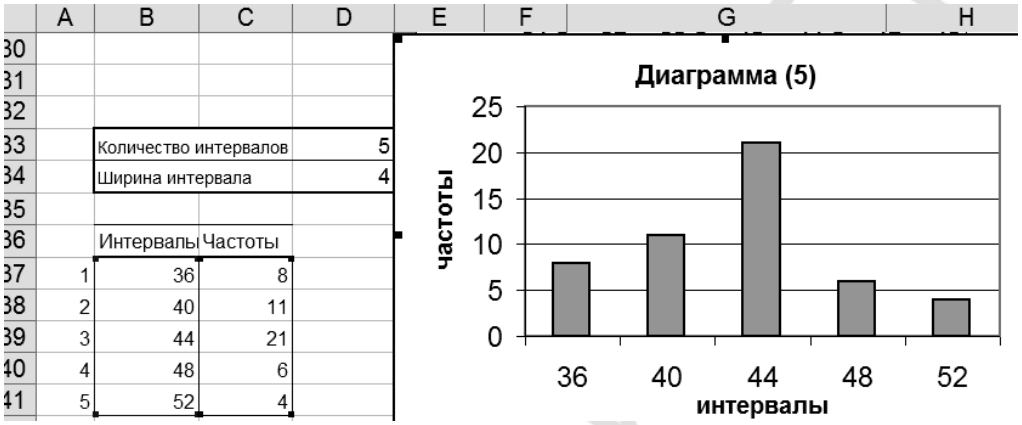
– выделите мышью диапазон ячеек **B37:B41**;

– нажмите кнопку разворачивания , подтвердите изменения, нажав **ОК**;

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

– постройте и отредактируйте гистограмму для нового варианта, задав заголовок **Гистограмма (5)**.

Результат ваших действий изображен на рис. 43.



Количество интервалов			5
Ширина интервала			4
	Интервалы	Частоты	
	1	36	8
	2	40	11
	3	44	21
	4	48	6
	5	52	4

Рис. 43

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 полигона частот, связывающего частоту попадания в интервал с правой границей каждого интервала.

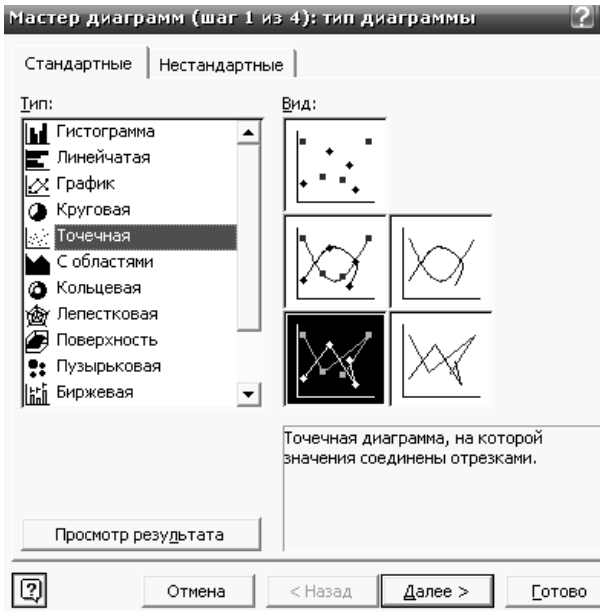


Рис. 44

1. Для построения полигона частот с применением мастера диаграмм выполните следующие действия:

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

– щелкните мышью кнопку **Мастер диаграмм** ;

– в окне диалога (рис. 44) на вкладке **Стандартные** в поле *Тип* выберите вариант **Точечная**;

– в поле **Вид** выберите **Точечная, на которой значения соединены отрезками**, и нажмите кнопку **Далее**;

– в появившемся окне на вкладке **Диапазон данных** включите переключатель **Ряды в: столбцах** 

, нажмите кнопку **Далее**;

– в следующем окне на вкладке **Заголовки** (рис. 45) напечатайте в поле *Название диаграммы* текст **Полигон частот**, в поле *Ось X (категорий)* — текст **концентрация альбумина, г**; в поле *Ось Y (значений)* — текст **частоты**;

Рис. 45

– на вкладке **Линии сетки** установите флажки ***основные линии*** в разделах *Ось X* и *Ось Y*;

– на вкладке **Легенда** снимите флажок *Добавить легенду*, нажмите кнопку **Далее>**;

– в появившемся окне выбора места расположения диаграммы (рис. 46) включите переключатель **имеющемся** и нажмите кнопку **Готово**.

2. Чтобы изменить формат горизонтальной оси:

– наведите указатель мыши на горизонтальную ось, нажмите правую кнопку мыши для вызова контекстного меню (рис. 47);

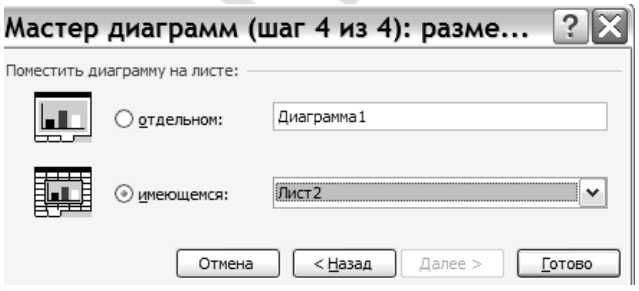


Рис. 46

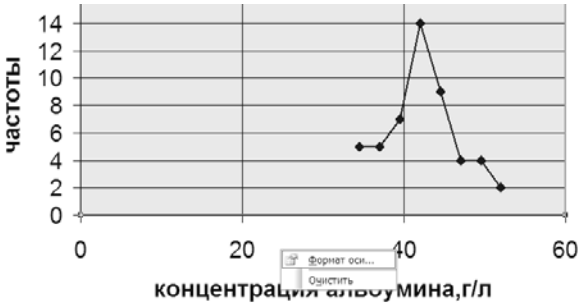


Рис. 47

– выберите в контекстном меню команду **Формат оси**;

– в открывшемся окне диалога **Формат оси** (рис. 48) на вкладке *Шкала* выберите следующие параметры: в поле минимальное значение введите «32», максимальное значение «52», цена основных делений «2», ось Y пересекает в значении «32». Нажмите **ОК**.

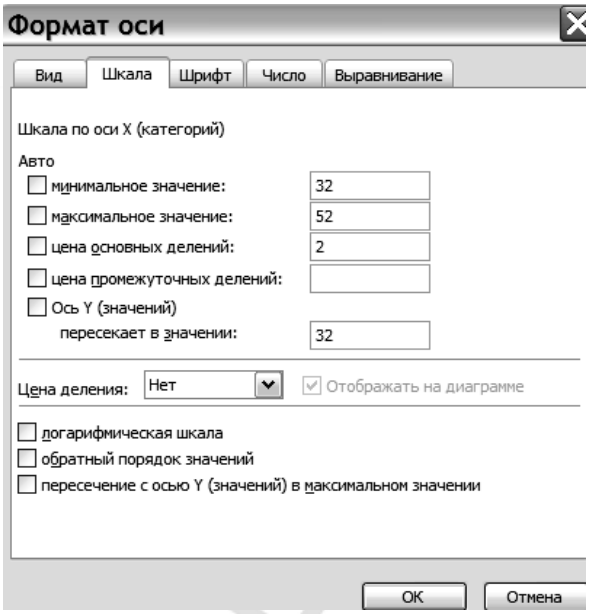


Рис. 48

3. Снимите заливку области построения графика, для чего:

– выделите область построения (сетку) щелчком правой кнопки мыши;

– в контекстном меню выберите **Формат области построения**;

– на вкладке **Вид** в области *Заливка* над палитрой установите переключатель прозрачная.

Результат ваших действий показан на рис. 49.

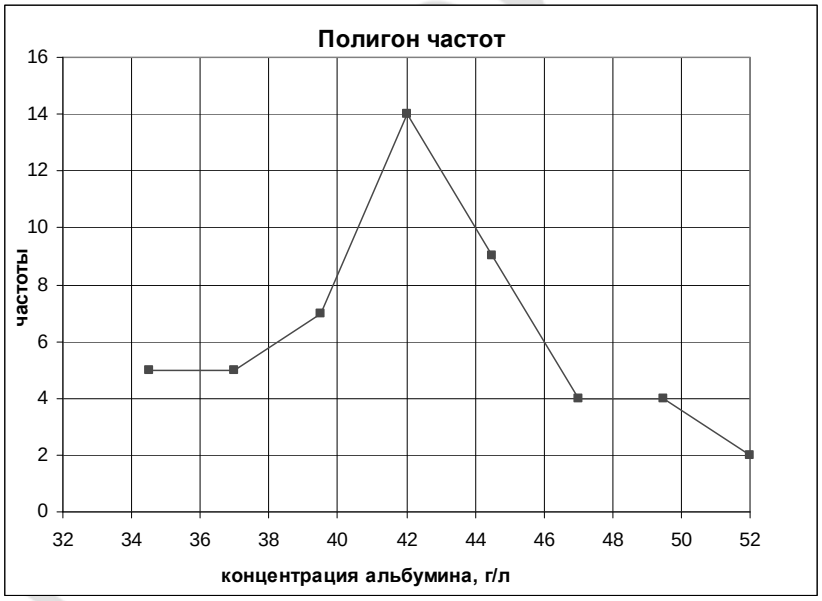


Рис. 49

Следует подчеркнуть, что результаты, получаемые с помощью встроенных статистических функций **Мастера функций** и графики, построенные с помощью **Мастера диаграмм**, 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 одномоментно. Такой метод расчета требует **обязательного введения данных 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**, выберите в меню **Правка** команду **Копировать**;

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

4. Данные задачи необходимо ввести в таблицу в виде столбца, для этого:

– на листе **Пакет анализа** выделите мышью диапазон **B2:B11**, выберите в меню **Правка** команду **Вырезать**;

– установите курсор в ячейку **A12**, выберите в меню **Правка** команду **Вставить**;

– далее выделите мышью диапазон **C2:C11**, выберите в меню **Правка** команду **Вырезать**;

– установите курсор в ячейку **A22**, выберите в меню **Правка** команду **Вставить**;

– описанным образом вырежьте по очереди и скопируйте диапазоны с данными **D2:D12, E2:E12**.

5. Запустите пакет анализа:

– выберите в меню **Сервис** команду **Анализ данных** (рис. 50);

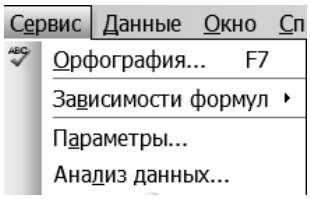


Рис. 50

– если эта команда отсутствует в списке, чтобы добавить ее, в меню **Сервис** выберите **Настройки**, а в открывшемся окне выделите мышью **Пакет анализа**;

– подтвердите выбор, нажав **ОК**;

– вернитесь к меню **Сервис** и выберите щелчком команду **Анализ данных**.

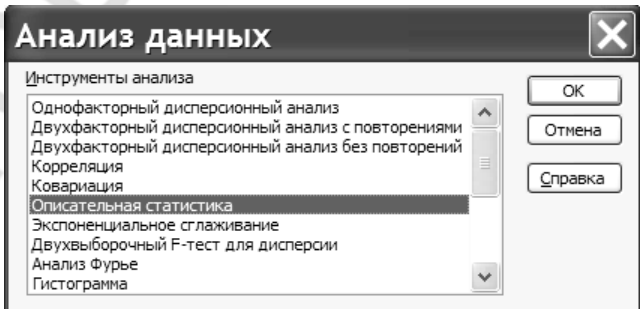


Рис. 51

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

7. Укажите в окне **Описательная статистика** (рис. 52) 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 **G2**;

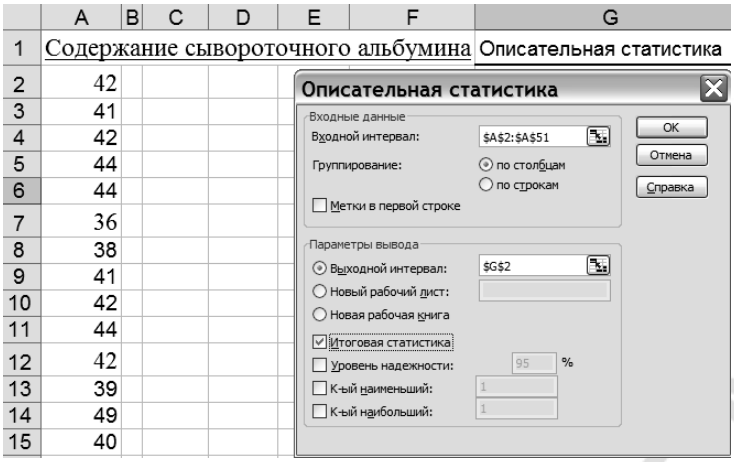


Рис. 52

– 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**.

Результаты расчетов приведены на рис. 53.

	A	G	H
1	Содержание сывороточного альбумина	Описательная статистика	
2	42	*Столбец1*	
3	41		
4	42	Среднее	41,38
5	44	Стандартная ошибка	0,67121
6	44	Медиана	42
7	36	Мода	42
8	38	Стандартное отклонение	4,746169
9	41	Дисперсия выборки	22,52612
10	42	Эксцесс	-0,17971
11	44	Асимметричность	0,027429
12	42	Интервал	20
13	39	Минимум	32
14	49	Максимум	52
15	40	Сумма	2069
16	45	Счет	50
17	32		
18	34		

Рис. 53

Сопоставьте результаты, полученные с помощью Пакета анализа на листе **Пакет анализа**, с результатами расчетов с помощью **Мастера функций** на листе **Описательная статистика** (рис. 23 и 53).

3.4.2. Построение гистограммы с помощью Пакета анализа

Выполните следующие действия:

1. Перейдите на новый лист, щелкнув на имени **Лист2**.
2. Запустите пакет анализа: выберите в меню **Сервис** команду **Анализ данных**.

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

4. 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**);

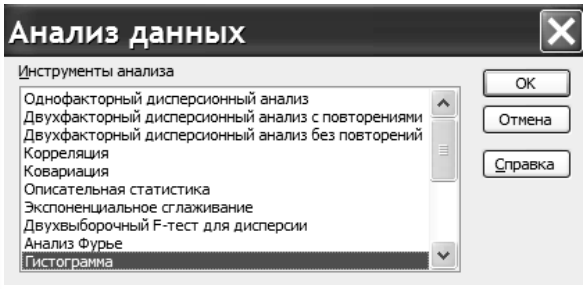


Рис. 54

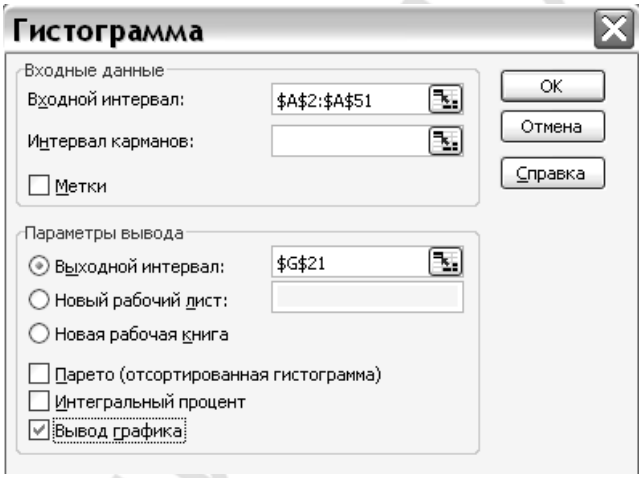


Рис. 55

– **выходной интервал** — щелкните мышью на ячейку в электронной таблице, начиная с которой будут выводиться результаты анализа, в нашем примере это ячейка **G21**;

– установите флажок **Вывод графика** для автоматического создания диаграммы на листе, содержащем выходной диапазон, подтвердите параметры, нажав **ОК**.

Результат отображен на рис. 56.

Карман	*Частота*
32	2
34,85714286	3
37,71428571	5
40,57142857	9
43,42857143	17
46,28571429	6
49,14285714	6
Еще	2

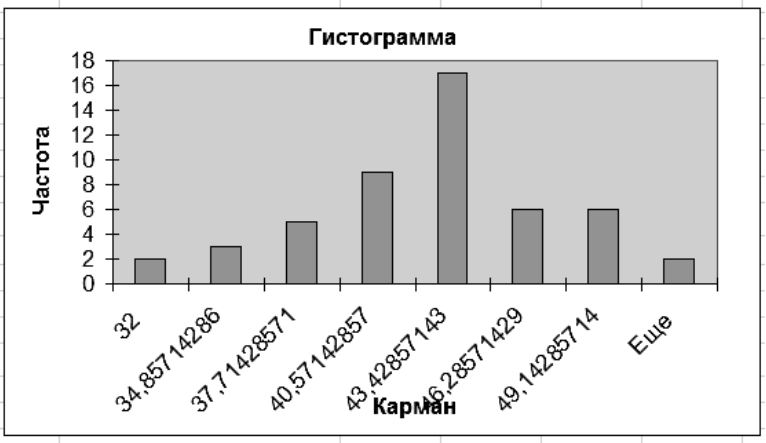


Рис. 56

Сравните гистограмму, построенную вручную на листе **Описательная статистика** (рис. 34), и гистограмму, построенную с помощью Пакета анализа, на листе **Пакет анализа** (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

Определите следующие статистические характеристики для этого показателя: выборочные среднее, медиану, моду, 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 среднего выборочного.
2. Расчет доверительных интервалов для дисперсии.
3. Расчет доверительных интервалов для стандартного отклонения.

Задача. Используя данные задачи, приведенной в подразд. 3.2, постройте доверительные интервалы для среднего, стандартного отклонения и дисперсии, 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. Запустите пакет Excel. В результате на экране появится окно программы, а в нем окно документа **Книга1**.

2. Откройте сохраненную на предыдущих занятиях книгу под именем **Статистика.xls** в своей рабочей папке, для этого:

– в меню **Файл** выберите команду **Открыть**;

– в появившемся окне *Открытие документа* в поле **Папка** выберите рабочую папку, куда был сохранен документ;

– найдите и выделите щелчком имя файла *Статистика*, нажмите кнопку **Открыть.**

3. Присвойте чистому листу **Лист3** новое имя **Доверит интервалы**:

– подведите курсор к имени листа, щелкните правой кнопкой мыши;

– в контекстном меню выберите **Переименовать**, введите новое имя.

4. Скопируйте на новый лист таблицу с исходными данными задачи «Содержание сывороточного альбумина» и вычисленными ранее числовыми характеристиками (рис. 57), для этого:

	A	B	C	D	E	F	G	H
1	Содержание сывороточного альбумина						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		Объём выборки	50
8	39	35	43	44	47		Асимметрия	0,027429375
9	40	39	42	41	46		Эксцесс	-0,17971336
10	37	49	41	39	43		Дисперсия	22,52612245
11	42	47	48	51	52		Стандартное отклонение	4,746169239
12							Вариационный размах	20
13							Коэффициент вариации	0,114697178
14							Стандартная ошибка	0,671209691

Рис. 57

– выделите диапазон ячеек **A1:H14**;

– в меню **Правка** выберите команду **Копировать**;

– перейдите на новый лист, установите курсор в ячейку **A1**;

– выберите в меню **Правка** команду **Вставить**.

3.5.2. Редактирование таблицы данных

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

1. На новом листе, в скопированном блоке, удалите ячейки с расчетами, невостребованными в данной задаче (выделенные диапазоны **G3:H6**; **G8:H9**; **G12:H13**):

– выделите мышью диапазон ячеек **G3:H6**;

– выберите в меню **Правка** команду **Удалить**;

– в окне диалога *Удаление ячеек* (рис. 58) поставьте переключатель в позицию *ячейки со сдвигом вверх*.

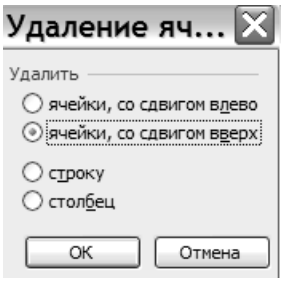


Рис. 58

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

3. Дополните таблицу данных — введите названия величин с **G7** по **G8** и числовое значение одной из них в **H7** как показано на рис. 60:

– выделите ячейку **G7** щелчком мыши; введите в нее текст «уровень значимости», завершите ввод, нажав клавишу **Enter**;

– выделите ячейку **H7**, введите в нее число «0,05», завершите ввод, нажав клавишу **Enter**;

– выделите ячейку **G8**, введите в нее текст «Доверительный интервал», зафиксируйте результат нажатием клавиши **Enter**.

	A	B	C	D	E	F	G	H
1	Содержание сывороточного						Описательная статистика	
2	42	41	42	44	44		Среднее	41,38
3	36	38	41	42	44		Объём выборки	50
4	42	39	49	40	45		Дисперсия	22,52612
5	32	34	43	37	39		Стандартное отклонение	4,746169
6	41	39	48	42	43		Стандартная ошибка	0,67121
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

	G	H
1	Описательная статистика	
2	Среднее	41,38
3	Объём выборки	50
4	Дисперсия	22,52612
5	Стандартное отклонение	4,746169
6	Стандартная ошибка	0,67121
7	уровень значимости	0,05
8	Доверительный интервал	

Рис. 60

3.5.3. РАСЧЕТ ДОВЕРИТЕЛЬНОГО ИНТЕРВАЛА ДЛЯ СРЕДНЕГО ВЫБОРОЧНОГО

1. Для расчета доверительного интервала выполните следующие, уже знакомые вам, действия:

– установите курсор в ячейку **H8**;

– нажмите кнопку **Вставка функции**;

– в окне диалога **Мастер функций** в поле *Категория* выберите **Статистические**;

– в поле *Выберите функцию*, листая список названий функций, найдите и выделите щелчком функцию **ДОВЕРИТ**;

– подтвердите выбор, нажав **ОК**.

На экране появится окно **Аргументы функции** (рис. 61), которое требует указания на ячейки, содержащие предварительно рассчитанные стандартное отклонение и объем (размер) выборки.

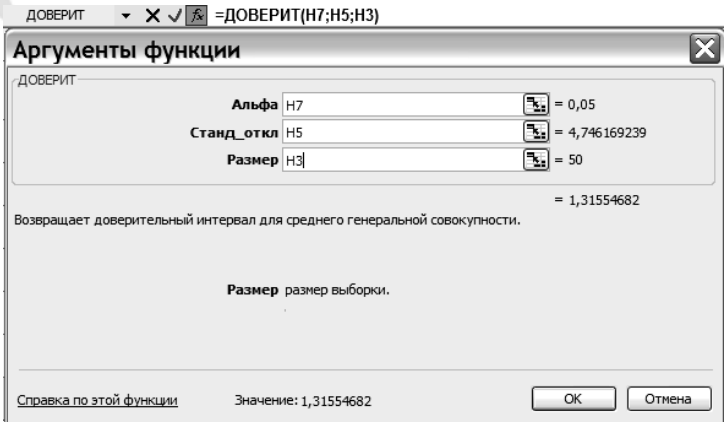


Рис. 61

2. Заполните поля диалогового окна **Аргументы функции**:

– в поле **Альфа** с помощью клавиатуры введите заданный уровень значимости **0,05** или укажите адрес ячейки **H7**;

– в поле **Станд_откл** введите адрес ячейки **H5**, содержащей стандартное отклонение, вручную или выделив эту ячейку в таблице щелчком мыши;

– в поле **Размер** укажите объем выборки: выделите в таблице ячейку **H3** щелчком мыши, или введите адрес этой ячейки вручную;

– подтвердите, нажав **ОК**.

Полученный числовой результат отобразится в ячейке **H8**.

Доверительный интервал	1,31554682

3.5.4. Расчет границ доверительного интервала для среднего значения

1. Заполните диапазон **J1:J3** текстовыми значениями по образцу, отображенному на рис. 62, для чего:

– в ячейку **J1** введите текст «Доверительный интервал для среднего», завершите ввод, нажав клавишу **Enter**;

– в ячейку **J2** введите текст «нижняя граница», нажмите клавишу **Enter**;

– в ячейку **J3** введите текст «верхняя граница», завершите ввод, нажав клавишу **Enter**.

	G	H	I	J	K
1	Описательная статистика			Доверительный интервал для среднего	
2	Среднее	41,38		нижняя граница	
3	Объём 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. 62

2. Увеличьте ширину столбца **J**:

– установите курсор в ячейку **J3** щелчком мыши;

– дважды щелкните правую границу заголовка столбца **J**.

3. В ячейку **K2** введите формулу **=H2–H8** для расчета нижней границы доверительного интервала:

– установите курсор в ячейку **K2**;

– нажмите клавишу = (равно);

– укажите щелчком мыши ячейку **H2**, где хранится уменьшаемое, при этом адрес этой ячейки автоматически заносится в формулу;

– нажмите клавишу с символом операции – (минус);

– укажите ячейку **H8**, где хранится вычитаемое, при этом адрес этой ячейки также заносится в формулу;

– для получения результата расчета нажмите **Enter**.

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

– установите курсор в ячейку **К3**, нажмите на клавиатуре клавишу = (равно);

– укажите щелчком мыши ячейку **Н2** с первым слагаемым, при этом адрес этой ячейки автоматически заносится в формулу;

– нажмите клавишу с символом операции + (плюс);

– укажите ячейку **Н8** со вторым слагаемым, адрес этой ячейки также заносится в формулу;

– для получения результата расчета нажмите **Enter**.

J	K
Доверительный интервал для среднего	
нижняя граница	40,064
верхняя граница	42,696

Рис. 63

Вычисленные границы доверительного интервала для среднего отображены на рис. 63.

Значение сывороточного альбумина с вероятностью 95 % лежит в интервале $40{,}064 < x < 42{,}696$.

Самостоятельно вычислите в столбце L границы нового доверительного интервала, приведите результат в виде $M_г(X) = \bar{x}_в \pm S/\sqrt{n}$ (при t = 1) и убедитесь, что границы доверительного интервала сужаются $40{,}7 < x < 42{,}1$. Приведенный интервал будет накрывать истинное значение величины с вероятностью 68 %.

3.5.5. Расчет доверительного интервала для дисперсии

1. Заполните названиями вычисляемых параметров ячейки **J5:J9** (согласно рис. 64).

– в ячейку **J5** введите текст «Доверительный интервал для дисперсии», завершите ввод, нажав клавишу **Enter**;

– в ячейку **J6** введите текст «коэффициент χ^2 нижнее», нажмите клавишу **Enter**;

– в ячейку **J7** введите текст «коэффициент χ^2 верхнее», нажмите клавишу **Enter**;

	J	K
1	Доверительный интервал для среднего	
2	нижняя граница	=Н2-Н6
3	верхняя граница	=Н2+Н6
4		
5	Доверительный интервал для дисперсии	
6	коэффициент χ2 нижнее	=ХИ2ОБР(0,05/2;Н3-1)
7	коэффициент χ2 верхнее	=ХИ2ОБР(1-0,05/2;Н3-1)
8	нижняя граница	=(Н3-1)*Н4/К6
9	верхняя граница	=(Н3-1)*Н4/К7
10		
11	Доверительный интервал для ст. отклоне	
12	нижняя граница	=КОРЕНЬ(К8)
13	верхняя граница	=КОРЕНЬ(К9)

Рис. 64

– в ячейку **J8** введите текст «нижняя граница», нажмите клавишу **Enter**;

– в ячейку **J9** введите текст «верхняя граница», завершите ввод, нажав клавишу **Enter**.

2. Вычислите с помощью статистической функции **ХИ2ОБР** коэффициент χ^2 нижнее, для чего:

– установите курсор в ячейку **К6**;

– нажмите кнопку f_x **Вставка функции**;

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

– в поле *Выберите Функцию*, листая список названий функций, найдите и выделите щелчком функцию **ХИ2ОБР**;

– подтвердите выбор, нажав **ОК**;

– в появившемся окне **Аргументы функции** (рис. 65) в поле *Вероятность* вручную введите **0,05/2** ($^1/_2$ уровня значимости);

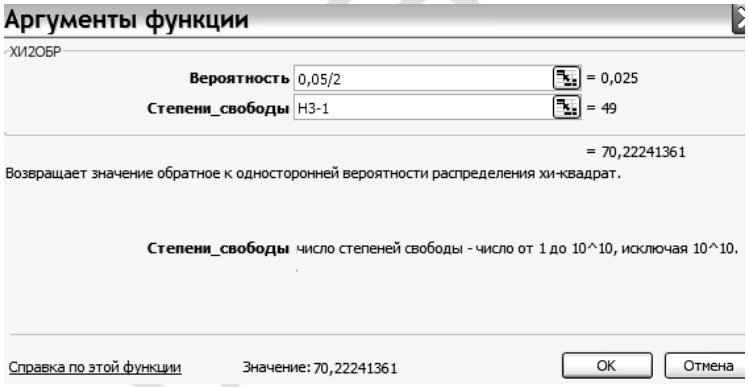


Рис. 65

– в поле *Степени_свободы* введите **Н3-1** (объем выборки — 1), щелкнув в ячейке **Н3**, где вычислен объем выборки и введя с клавиатуры **–1** (минус один); нажмите **ОК** в окне.

3. Вычислите с помощью статистической функции **ХИ2ОБР** коэффициент χ^2 верхнее, для чего:

– установите курсор в ячейку **К7**, нажмите кнопку f_x **Вставка функции**;

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

– в поле *Выберите Функцию*, листая список названий функций, найдите и выделите щелчком функцию **ХИ2ОБР**, подтвердите выбор, нажав **ОК**;

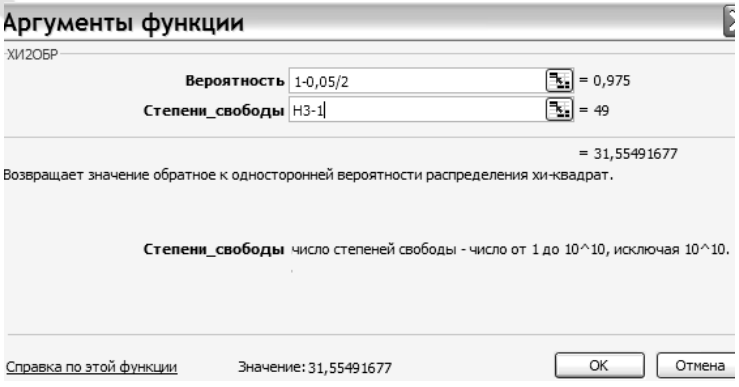


Рис. 66

– в появившемся окне **Аргументы функции** (рис. 66) в поле *Вероятность* вручную введите **1–0,05/2** ($1–^1/_2$ уровня значимости);

– в поле *Степени_ свободы* введите (объем выборки — 1), щелкнув в ячейке **Н3**, где вычислен объем выборки и введя с клавиатуры **–1** (минус один), нажмите **ОК** в окне.

4. Введите в ячейку **К8** формулу **=(Н3-1)*Н4/К6** для расчета нижней границы доверительного интервала (рис. 64):

– нажмите на клавиатуре клавишу = (равно); затем нажмите ((скобку);

– укажите щелчком мыши ячейку **Н3** (объем выборки), при этом адрес этой ячейки автоматически заносится в формулу;

– нажмите клавишу с символом операции – (минус), затем нажмите 1 и) (скобку), затем нажмите * (умножить);

– укажите щелчком мыши ячейку **Н4** со значением дисперсии, адрес этой ячейки также заносится в формулу;

– нажмите клавишу с символом операции / (деление),

– укажите щелчком мыши ячейку **К6** (коэффициент χ^2 нижнее)

– для получения результата расчета нажмите **Enter**.

5. Введите в ячейку **К9** формулу **=(Н3-1)*Н4/К7** для расчета верхней границы доверительного интервала:

– нажмите на клавиатуре клавишу = (равно); затем нажмите ((скобку);

– укажите щелчком мыши ячейку **Н3** (объем выборки), при этом адрес этой ячейки автоматически заносится в формулу;

– нажмите на клавиатуре клавишу с символом операции – (минус), затем нажмите 1 и) (скобку), затем нажмите * (умножить);

– укажите щелчком мыши ячейку **Н4** со значением дисперсии, адрес этой ячейки также заносится в формулу;

– нажмите клавишу с символом операции / (деление);

– укажите ячейку **К7** (коэффициент χ^2 верхнее);

– для получения результата расчета нажмите **Enter**.

	J	K	L
5	Доверительный интервал для дисперсии		
6	коэффициент χ2 нижнее	70,222	
7	коэффициент χ2 верхнее	31,555	
8	нижняя граница	15,718	
9	верхняя граница	34,98	

Рис. 67

Результат расчетов отображен на рис. 67.

Величина дисперсии для сывороточного альбумина с вероятностью 95 % лежит в интервале $15{,}72 < D < 39{,}98$.

3.5.6. Расчет доверительного интервала для стандартного отклонения

1. Заполните названиями вычисляемых параметров ячейки **J11: J13** по образцу на рис. 68:

– в ячейку **J11** введите текст «Доверительный интервал для стандартного отклонения», завершите ввод, нажав клавишу **Enter**;

	J	K
11	Доверительный интервал для ст. отклонени	
12	нижняя граница	=КОРЕНЬ(K8)
13	верхняя граница	=КОРЕНЬ(K9)

Рис. 68

– в ячейку **J12** введите текст «нижняя граница», нажмите клавишу **Enter**;

– в ячейку **J13** введите текст «верхняя граница», нажмите клавишу **Enter**.

2. Вычислите в ячейке **К12** с помощью математической функции **КОРЕНЬ** нижнюю границу доверительного интервала, для чего:

– установите курсор в ячейку **К12**, нажмите кнопку **Вставка функции**;

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

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

– в появившемся окне **Аргументы функции** (рис. 69) в поле *Число* укажите щелчком адрес ячейки **К8** (где вычислена нижняя граница интервала для дисперсии);

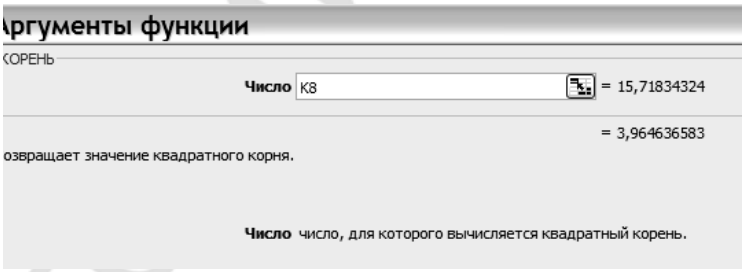


Рис. 69

– нажмите **ОК** в окне.

3. Таким же образом вычислите в ячейке **К13** с помощью математической функции **КОРЕНЬ** верхнюю границу доверительного интервала, для чего:

– установите курсор в ячейку **К13**, нажмите кнопку **Вставка функции**;

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

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

– в появившемся окне **Аргументы функции** в поле *Число* укажите щелчком адрес ячейки **К9** (где вычислена верхняя граница интервала для дисперсии), нажмите **ОК** в окне.

Результат вычислений отображен на рис. 70.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Содержание сывороточного						Описательная статистика			Доверительный интервал для среднего		
2	42	41	42	44	44		Среднее	41,38		нижняя граница	40,064	
3	36	38	41	42	44		Объём выборки	50		верхняя граница	42,696	
4	42	39	49	40	45		Дисперсия	22,52612				
5	32	34	43	37	39		Стандартное отклонение	4,746169		Доверительный интервал для дисперсии		
6	41	39	48	42	43		Стандартная ошибка	0,67121		коэффициент $\chi 2$ верхнее	70,222	
7	33	43	35	32	34		уровень значимости	0,05		коэффициент $\chi 2$ нижнее	31,555	
8	39	35	43	44	47		Доверительный интервал	1,315547		нижняя граница	15,718	
9	40	39	42	41	46					верхняя граница	34,98	
10	37	49	41	39	43							
11	42	47	48	51	52					Доверительный интервал для ст. отклонения		
12										нижняя граница	3,9646	
13										верхняя граница	5,9144	

Рис. 70

Значение стандартного отклонения для сывороточного альбумина с вероятностью 95 % лежит в интервале $3{,}96 < \sigma < 5{,}91$.

3.5.7. 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.xls** 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 (рис. 70).
4. Уменьшите объем выборки с 50 до 35, удалив данные из ячеек **A9:E11** как показано на рис. 71.
5. Пользуясь строкой формул, отредактируйте формулы для расчета числовых характеристик выборки, находящиеся в ячейках **H2:H5**, изменив в них адрес диапазона ячеек с данными для вычислений, для чего:

– установите курсор в ячейку **H2**, в строке формул, справа от кнопки f_x, отобразится введенная в эту ячейку формула **=СРЗНАЧ(A2:E11)**;

– установите курсор в строке формул и замените 11 на 8, чтобы формула получила следующий вид: **=СРЗНАЧ(A2:E8)**;

– таким же образом отредактируйте формулы, находящиеся в ячейках **H3:H5**.

6. Проследите, как во всех остальных формулах для расчета доверительных интервалов автоматически выполняется перерасчет.

Результат ваших действий отображен на рис. 71.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Содержание сывороточного						Описательная статистика			Доверительный интервал для среднего		
2	42	41	42	44	44		Среднее	40,34286		нижняя граница	38,872	
3	36	38	41	42	44		Объём выборки	35		верхняя граница	41,813	
4	42	39	49	40	45		Дисперсия	19,70252				
5	32	34	43	37	39		Стандартное отклонение	4,438752		Доверительный интервал для дисперсии		
6	41	39	48	42	43		Стандартная ошибка	0,750286		коэффициент $\chi 2$ верхнее	51,966	
7	33	43	35	32	34		уровень значимости	0,05		коэффициент $\chi 2$ нижнее	19,806	
8	39	35	43	44	47		Доверительный интервал	1,470534		нижняя граница	12,891	
9										верхняя граница	33,822	
10												
11										Доверительный интервал для ст. отклонения		
12										нижняя граница	3,5904	
13										верхняя граница	5,8157	

Рис. 71

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 выборочных значений от среднего, а затем проверьте выполнений следующих условий:

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

Рассмотрим последовательность необходимых действий.

1. Вставьте в книгу под именем **Статистика** новый лист следующим образом:
   – в строке меню выберите команду **Вставка**;
   – в раскрывшемся подменю выберите **Лист**;
   – дайте листу имя **Проверка на норм**.
2. Скопируйте на новый лист таблицу с исходными данными задачи «Содержание сывороточного альбумина» — диапазон ячеек **A1:E11**.
3. Введите в ячейки **G13:G20** следующие названия вычисляемых характеристик по образцу на рис. 72.

	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. 72

4. Рассчитайте в ячейке **H13** среднее значение, используя известную вам функцию **СРЗНАЧ**, для этого выполните следующие действия:

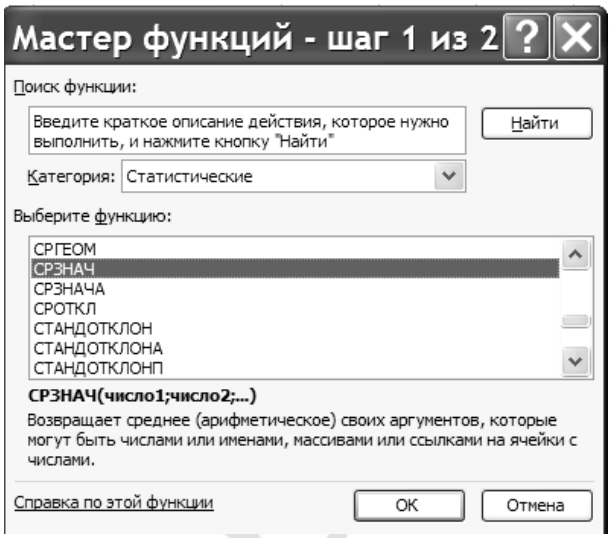


Рис. 73

– выделите щелчком мыши ячейку **H13**;

– нажмите кнопку **Вставка функции**;

– в появившемся окне **Мастер функций** (рис. 73) в поле *Категория* выберите **Статистические**;

– в поле *Выберите Функцию*, листая список названий функций, найдите и выделите функцию **СРЗНАЧ**;

– подтвердите выбор, нажав **ОК**;

– в появившемся окне *Аргументы функции* в поле **Число1** введите адрес диапазона ячеек с данными **A2:E11**, выделив этот диапазон в таблице мышью;

– подтвердите, нажав **ОК**.

	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. 74

5. Аналогично вычислите (рис. 74) в ячейке **H14** объем выборки, используя функцию **СЧЕТ**, а в ячейке **H15** стандартное отклонение, используя функцию **СТАНДОТКЛОН**, каждый раз указывая диапазон ячеек с данными **A2:E11**.

6. В ячейке **H16** вычислите значение **2S,** используя формулу **=2*H15**, для этого:

– установите курсор в ячейку **H16**;

– нажмите клавишу = (равно);

– введите с клавиатуры множитель 2, затем нажмите клавишу с символом операции * (умножить);

– укажите мышью ячейку **H15** (значение S), при этом адрес этой ячейки автоматически заносится в формулу;

– для получения результата расчета нажмите **Enter**.

7. В ячейке **H17** вычислите значение **3S,** используя формулу **=3*H15**:

– установите курсор в ячейку **H17**;

– нажмите клавишу = (равно);

– введите с клавиатуры множитель 3, затем нажмите клавишу с символом операции * (умножить);

– укажите щелчком мыши ячейку **H15** (значение S), при этом адрес этой ячейки автоматически заносится в формулу;

– для получения результата вычислений нажмите **Enter**.

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

9. Рассчитайте массив отклонений выборочных значений от среднего, для чего в ячейку **A14** введите функцию **=ABS(A2–$H$13)** как показано на (рис. 75):

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

Рис. 75

– установите курсор в ячейку **A14**;

– выберите в меню **Вставка** команду **Функция**. В категории *Математические* выберите функцию **ABS**;

– установите курсор в поле *Число* и введите ссылку на ячейку **A2**, введите знак –, а затем ссылку на ячейку **H13**;

– сделайте ссылку **H13** абсолютной, нажав функциональную клавишу **F4**;

– подтвердите, нажав **ОК**.

10. Скопируйте введенную в ячейку **A14** формулу вниз до **A23**, затем вправо до **E23**. Результат расчетов отображен на рис. 76.

11. Подсчитайте в ячейках **H18:H20** сколько отклонений составляет 68,3; 95,5 и 99,7 % от их общего количества (рис. 77):

– установите курсор в ячейку **H18**;

– нажмите клавишу = (равно);

– введите с клавиатуры множитель 0,683, затем нажмите клавишу с символом операции * (умножить);

– укажите щелчком мыши ячейку **H14** (объем выборки), при этом адрес этой ячейки автоматически заносится в формулу;

– для получения результата расчета нажмите **Enter**.

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

	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. 76

	G	H
18	68,3%	=0,683*H14
19	95,5%	=0,955*H14
20	99,7%	=0,997*H14

Рис. 77

Результат вычислений отображен на рис. 78.

12. Создайте таблицу условий проверки распределения на нормальность, для чего:

– введите в ячейки **G21:G24** текстовые значения по образцу на рис. 79;

– в ячейку **H21** введите текст «Выполнение условия»;

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

Рис. 78

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

Рис. 79

13. Введите в ячейки **H22:H24** функции для вычисления критерия выполнения условия (подсчет отклонений удовлетворяющих заданному условию) как на рис. 80:

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

Рис. 80

– в ячейку **H22** введите функцию **СЧЕТЕСЛИ,** нажав кнопку f_x **Вставка функции**;

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

– в поле *Выберите Функцию*, листая список названий функций, найдите и выделите щелчком функцию **СЧЕТЕСЛИ**.

14. В окне **Аргументы функции** (рис. 81):

– в поле *Диапазон* укажите ссылку на диапазон ячеек **A14:E23**;

– в поле *Критерий* введите **<4,75**;

– подтвердите выбор, нажав **ОК**.

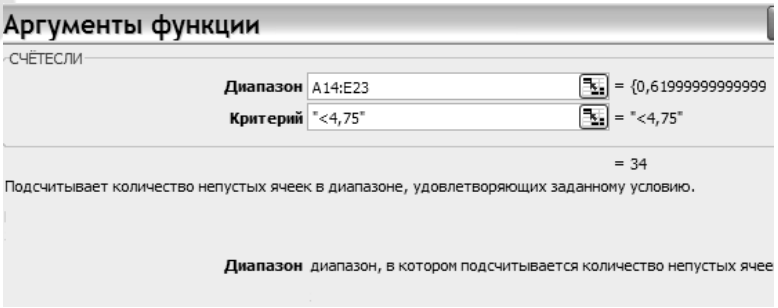


Рис. 81

15. Аналогичным образом введите формулы в ячейках **H23** и **H24** по образцу на рис. 80.

Результат ваших действий отображен на рис. 82.

	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		объём 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		Условие	Выполнение условия	
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. 82

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

Основные вопросы:

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

Задача № 1. Случайная величина X — сердечный индекс (л/мин·м^2). Известно, что в популяции здоровых людей его среднее значение $M(X)$ =

[1] В табличном процессоре Excel термин критерий заменяется словом тест.

А = 3,5 л/мин·м^2. По данным выборки из 112 пациентов, находящихся в критическом состоянии, среднее выборочное значение $\bar{x}_в = 2{,}45(\frac{л}{мин \cdot м^2})$, стандартное отклонение равно 1,32 л/мин·м^2. Известно, что у критически больных пациентов кровообращение замедленно. Нужно установить, будет ли $M_1(x)$ критических больных меньше чем 3,5 л/мин·м^2.

Гипотеза H_0: $M_1(X) = A = 3{,}5$ л/мин·м^2.

Альтернативная гипотеза H_1: $M_1(X) < A = 3{,}5$ л/мин·м^2 (левосторонняя критическая область). Уровень значимости α = 0,05.

	A	B
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	t критическое	

Рис. 83

1. Создайте новый файл, используя в меню **Файл** команду **Создать**.

2. Сохраните новую книгу под именем **Гипотезы** в рабочую папку.

3. Переименуйте **Лист1** в **О равенстве А**.

4. Введите данные задачи № 1 согласно образцу на рис. 83.

5. Выполните в ячейке **B7** расчет t-статистики (рис. 84) по формуле $t = \frac{(\overline{X}_в - A)\sqrt{n}}{S}$, для чего:

	B
7	=(B4-B2)*КОРЕНЬ(B3)/B5
8	=СТЬЮДРАСПОБР(0,1;B3-1)

Рис. 84

– установите курсор в ячейку **B7**;

– нажмите клавишу = (равно);

– откройте скобки, нажав клавишу ((скобка);

– укажите щелчком мыши на ячейку **B4**, содержащую среднее выборочное;

– нажмите клавишу – (минус);

– укажите щелчком мыши на ячейку **B2**, содержащую предполагаемое среднее;

– закройте скобки, нажав клавишу) (скобка);

– нажмите клавишу * (умножить);

– вставьте функцию **КОРЕНЬ**, нажав кнопку Вставка функции f_x;

– укажите щелчком мыши на ячейку **B3**, содержащую объем выборки;

– нажмите клавишу / (деление);

– укажите щелчком мыши на ячейку **B5**, содержащую стандартное отклонение;

– для получения результата нажмите **Enter**.

6. Вычислите в ячейке **B8** с помощью статистической функции значение t-критическое, для чего:

– вставьте функцию **СТЬЮДРАСПОБР**, нажав кнопку Вставка функции f_x;

– в окне диалога **Аргументы функции** (рис. 85) в поле *Вероятность* введите с клавиатуры **2·0,05**[1] (так как критическая область двусторонняя);

– в поле *Степени_свободы* введите, щелкнув мышью в ячейке **B3**, которая содержит объем выборки; а затем **–1** (минус один) — с клавиатуры.

7. Результат ваших действий отображен на рис. 86.

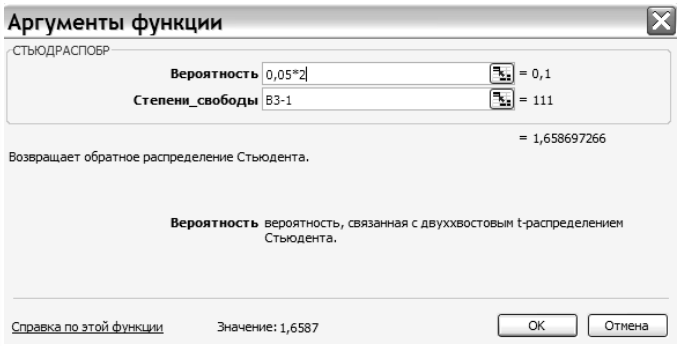


Рис. 85

	A	B
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

Рис. 86

Сравните t наблюдаемое и t критическое. Так как критическая область левосторонняя, то t критическое должно быть отрицательным, т. е. равно –1,6587.

Так как t наблюдаемое меньше t критического (–8,42 < –1,66), значит гипотеза H_0 отвергается, т. е. среднее генеральной совокупности критических больных значимо отличается от того же показателя для популяции здоровых людей.

8. Введите в ячейку **A9** текст результата анализа: «t наблюдаемое < t критического, гипотеза $M_1(X) = 3{,}5$ отвергается».

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

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

2. В ячейку **B10** вставьте функцию для расчета этой характеристики (рис. 87):

	A	B
10	доверительный интервал	=ДОВЕРИТ(0,05;B5;B3)

Рис. 87

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

[1] 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 СТЬЮДРАСПОБР(0,05;10). Одностороннее значение для той же значимости и того же числа степеней свободы может быть вычислено формулой СТЬЮДРАСПОБР(2*0,05;10).

– в окне диалога **Аргументы функции** (рис. 88) в поле *Альфа* введите с клавиатуры уровень значимости **0,05**;

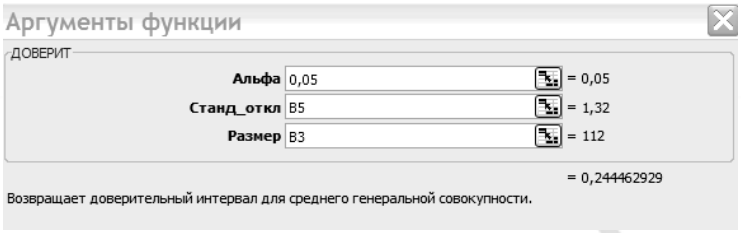


Рис. 88

– 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**.

Результат всех расчетов задач № 1 и 2 отображен на рис. 89.

	A	B
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. 89

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

Таким образом, с вероятностью 95 % интервал (2,45 ± 0,24) л/мин·м^2 содержит истинное среднее значение сердечного индекса критически больных пациентов. Предполагаемое A = 3,5 л/мин·м^2, 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: $X_в$ = 22 с, S = 3 с. Можно ли, по данным результатам, при уровне значимости 0,02 отклонить 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 α = 0,02.

1. Введите данные задачи № 3 согласно образцу на рис. 90.
2. Выполните в ячейке **B19** расчет t-статистики (рис. 91) по формуле $T = \dfrac{(\overline{X}_в - A)\sqrt{n}}{S}$, для чего:

– установите курсор в ячейку **B19**;

– вычислите разность (**B16-B14**);

– нажмите клавишу * (умножить);
– вставьте функцию **КОРЕНЬ**, нажав кнопку Вставка функции f_x;
– укажите щелчком мыши на ячейку **B15**, содержащую объем выборки;
– нажмите клавишу / (деление);
– укажите на ячейку **B17**, содержащую стандартное отклонение;
– для получения результата нажмите **Enter**.

	A	B
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. 90

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

Рис. 91

3. Вычислите, с помощью статистической функции, значение t критическое, для чего:
– в ячейку **B19** вставьте функцию **СТЬЮДРАСПОБР**, нажав кнопку Вставка функции f_x;
– в окне диалога **Аргументы функции** (рис. 92) в поле *Вероятность* введите с клавиатуры **0,02**[1] (так как критическая область двусторонняя);
– в поле *Степени_свободы* введите, щелкнув мышью в ячейке **B15**, содержащей объем выборки; затем**–1** (минус один) — с клавиатуры.
4. Результат ваших действий отображен на рис. 93.

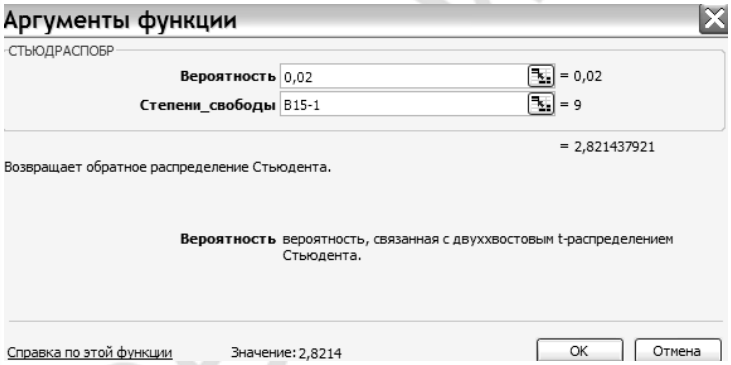


Рис. 92

	A	B
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. 93

Сравните t наблюдаемое и t критическое. Так как t наблюдаемое по модулю меньше t критического ($|2{,}11| < 2{,}82$), гипотеза H_0 принимается —

[1] В поле *Вероятность* вводят вероятность, соответствующую двустороннему распределению Стьюдента.

$M(X)$ = 20 с не противоречит данным эксперимента и нет оснований менять установленную норму.

5. Введите в ячейку **A21** результат анализа в виде текста: «t наблюдаемое < t критического, гипотеза H_0: $M(X)$ = 20 с принимается».

3.7.2. Задачи для самостоятельного выполнения

Задача № 1. В эксперименте исследуется изменение размера карциномы Герена под влиянием магнитного поля низкой частоты. Исследуемая выборка состоит из 8 объектов. Среднее выборочное значение $\overline{X_в}$ = 0,124 см. Дисперсия D = 0,013 см^2. Необходимо проверить, действительно ли переменное магнитное поле приводит к уменьшению карциномы.

Проверьте гипотезу H_0: $M(X)$ = 0,08 см против конкурирующей гипотезы H_1: $M(X)$ < 0,08 см, с уровнем значимости α = 0,05.

Задача № 2. 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 α = 0,05.

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

Задача № 3. 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 α = 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 (1/ч). Исходные данные приведены в таблице:

– 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.

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(контр) = M(1-й день, опыт).
2. M(контр) = M(9-й день, опыт).

Альтернативные гипотезы H_1:

1. M(контр) ≠ M(1-й день, опыт).
2. M(контр) ≠ M(9-й день, опыт).

Контрольная группа	Опытная группа		
	1-й день	3-й день	9-й день
2	28	15	5
5	35	13	*3*
3	40	19	8
0	25	5	3
1	33	18	7
5	42	18	8
3	19	5	4
2	21	10	5
8	28	16	2
1	31	15	2
0			
6			
4			
2			
7			

Если модуль t наблюдаемого < t критического, а p > α, 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. 94.

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. 95) выберите, выделив щелчком мыши, **Двухвыборочный t-тест с одинаковыми дисперсиями**;

	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. 94

– в появившемся окне диалога (рис. 96) в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания и выделите мышью диапазон ячеек с данными **A2:A16**, нажмите кнопку разворачивания;

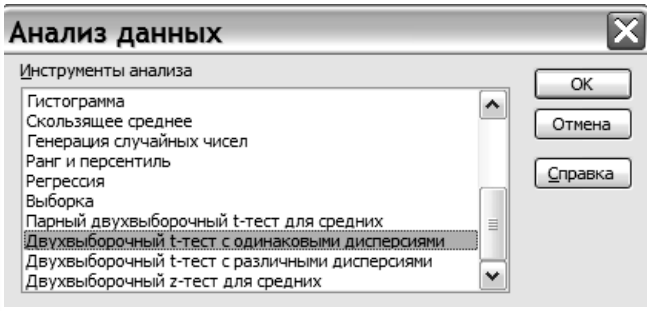


Рис. 95

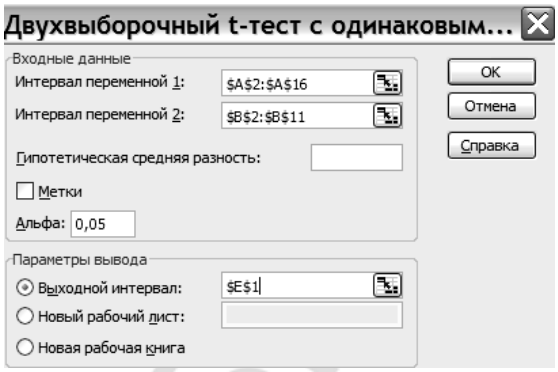


Рис. 96

– 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. 97.

	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. 97

Проанализируем результат: t-статистика (наблюдаемое) по модулю больше t критического (12,91 > 2,06), а P меньше α ($5{,}06 \cdot 10^{-12} << 0{,}05$), по-

этому гипотеза H_0 отвергается и принимается гипотеза H_1: M(контр) ≠ M(1-й день, опыт), различие выборочных средних также значимо.

6. Введите в ячейку **E15** текст результата анализа: «t-статистика > t критического, а P << α, гипотеза о равенстве средних отвергается».

Описанным выше способом выполните процедуру проверки гипотезы о равенстве средних, сравнив **контрольную группу** и **экспериментальную в 9-й день**.

1. Запустите процедуру проверки, для чего:

– в меню **Сервис** выберите **Анализ данных**;

– в окне **Анализ данных** (рис. 95) 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. 98.

	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. 98

Проанализируем результат: t-статистика (наблюдаемое) 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(контр) = M(9-й день, опыт) принимается. Выборочные средние различаются незначимо.

4. Введите в ячейку **E31** текст результата анализа: «t-статистика < t критического, а P > α, 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 одинаковыми 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.

3.7.5. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей с использованием команды Анализ данных

В задаче, данной в 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 α = 0,05, Гипотеза H_0: D(контр) = D(9-й день, опыт). Альтернативная гипотеза H_1: D(контр) > D(9-й день, опыт).

Если F наблюдаемое < F критического и p > α, 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 чего:

– в меню **Сервис** выберите **Анализ данных**;

– в окне **Анализ данных** (рис. 99) выберите, выделив щелчком мыши, **Двухвыборочный F-тест для дисперсии**;

– в появившемся окне диалога (рис. 100) в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **A2:A16**, нажмите кнопку разворачивания ;

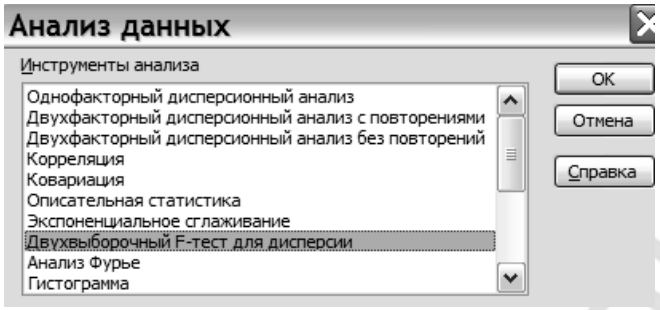


Рис. 99

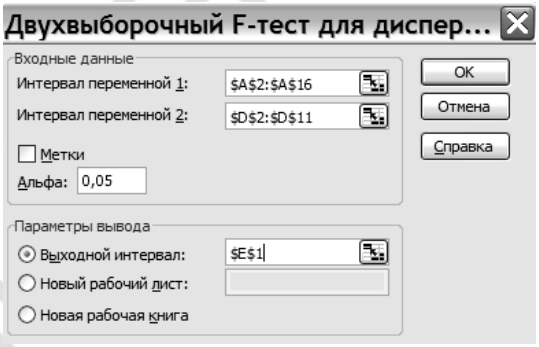


Рис. 100

– 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 **E1**, нажмите кнопку разворачивания ;

– 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. 101.

Проанализируем результат: F-статистика (наблюдаемое) меньше F-критического (1,16 < 3,03), а P больше α (0,2 > 0,05), поэтому гипотеза о равенстве дисперсий принимается (дисперсии двух генеральных совокупностей равны, различие генеральных дисперсий не значимо).

7. Введите в ячейку **E11** текст результата анализа: «F наблюдаемое < F критического, а P > α, гипотеза о равенстве дисперсий принимается».

	E	F	G
1	Двухвыборочный F-тест для дисперсии		
2	контроль/ 9 день		
3		*Переменная 1*	*Переменная 2*
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. Fрасч<F кр, P>α	гипотеза принимается	

Рис. 101

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 α = 0,05, проверьте нулевую гипотезу H_0: D(1-й день, опыт) = D(3-й день, опыт) о равенстве генеральных дисперсий, при конкурирующей гипотезе H_1: D(1-й день, опыт) > D(3-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).

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.

1-й день	3-й день	9-й день
28	15	5
35	13	3
40	19	8
25	5	3
33	18	7
42	18	8
19	5	4
21	10	5
28	16	2
31	15	2

Гипотеза H_0: M(3-й день) = M(9-й день)

Альтернативная гипотеза H_1: M(3-й день) ≠ M(9-й день)

Если модуль t наблюдаемого < t критического, а p > α, 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 **A1:D16**, содержащий выборочные данные.
2. Перейдите на новый рабочий лист и вставьте скопированный диапазон.
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. 102) выберите, выделив щелчком мыши, **Парный двухвыборочный t-тест для средних**;

– в появившемся окне диалога (рис. 103) в группе **Входные данные** поле *Интервал переменной 1* нажмите кнопку сворачивания  и выделите мышью диапазон ячеек с данными **C2:C11**, нажмите кнопку разворачивания  ;

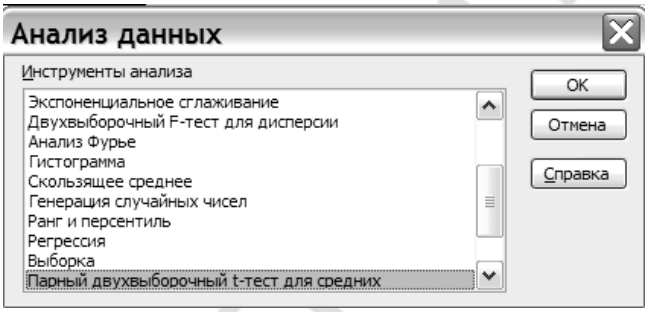


Рис. 102

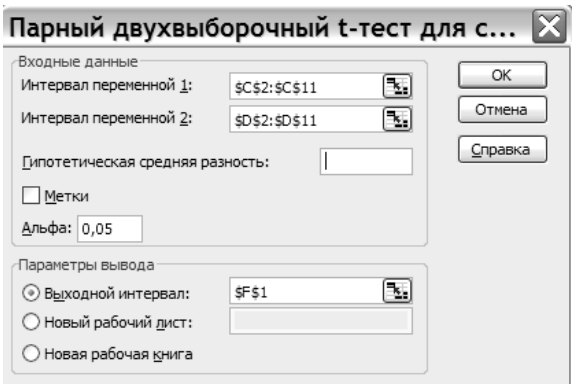


Рис. 103

– 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 *Гипотетическая средняя разность* введите **0** или оставьте поле пустым;

– в поле *Параметры вывода*, установите переключатель в позицию *Выходной интервал*;

– нажмите кнопку сворачивания и выделите мышью ячейку для вывода результатов **F1**, нажмите кнопку разворачивания;

– 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.

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

7. Введите в ячейку **E16** текст результата анализа: «t наблюдаемое > t критического, а P < α, гипотеза о равенстве средних отвергается».

Результат ваших действий отображен на рис. 104.

	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. 104

8. Сохраните результаты работы на диске.

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 36 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

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. 105. Для этого:
   - выделите ячейку **A1**;
   - введите в нее название исследуемого параметра «гемоглобин, %»;
   - завершите ввод, нажав клавишу **Enter**;
   - выделите ячейку **B1**;
   - введите в нее название второго исследуемого параметра «оседание эритроцитов»;
   - таким же образом заполните блок ячеек таблицы **A2:B15** числовыми данными.

	A	B
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. 105

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

4. Вычислите коэффициент корреляции, для чего:

– выделите щелчком мыши ячейку **D2**;

– нажмите кнопку **Вставка функции**;

– в появившемся окне **Мастер функций** (рис. 106) в поле **Категория** щелчком мыши выберите **Статистические**;

– в поле **Выберите Функцию**, листая список названий функций, найдите и выделите щелчком функцию **КОРРЕЛ**;

– подтвердите выбор, нажав **ОК**;

– в появившемся окне *Аргументы функции* (рис. 107) в поле **Массив1** вручную введите адрес первого диапазона ячеек с данными **A2:A15** или выделите этот диапазон в таблице мышью;

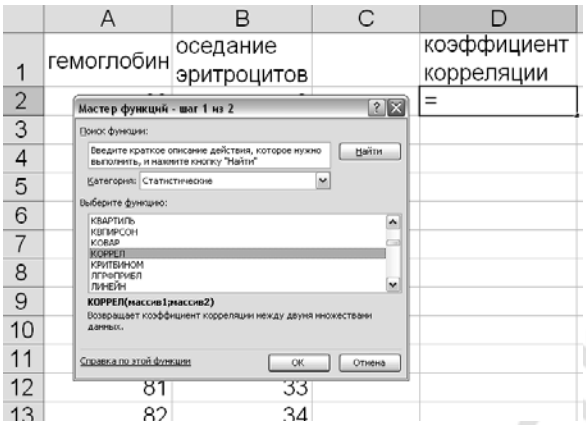


Рис. 106

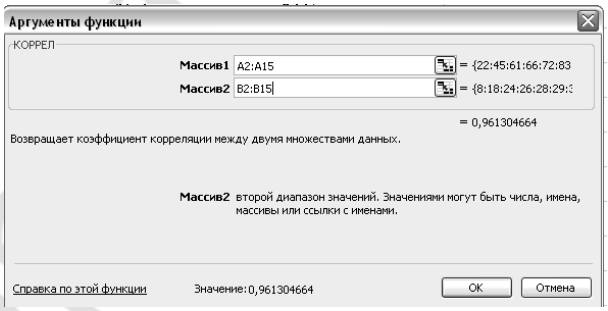


Рис. 107

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

Результат ваших действий отображен на рис. 108.

	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. 108

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 подразд. 3.8.1, в виде диаграммы рассеяния (корреляционного поля), иллюстрирующей связь между переменными. Постройте линию регрессии и получите уравнение регрессии.

1. Для построения диаграммы с применением мастера диаграмм выполните следующие действия:

– выделите диапазон ячеек с данными задачи **A2:B15**;

– щелкните мышью кнопку **Мастер диаграмм**  ;

– в окне диалога (рис. 109) на вкладке **Стандартные** в поле *Тип* выберите вариант *Точечная;*

– в поле **Вид** выберите, *Точечная позволяющая сравнить пары значений* и нажмите кнопку **Далее>**;

– в появившемся окне на вкладке **Диапазон данных** включите переключатель *Ряды в: столбцах* 

, нажмите кнопку **Далее>**;

– в следующем окне на вкладке **Заголовки** (рис. 110) напечатайте в полях *Название диаграммы* текст **Диаграмма рассеяния**, *Ось X (категорий)* — текст **гемоглобин, %**; *Ось Y (значений)* — текст **оседание эритроцитов, мм**;

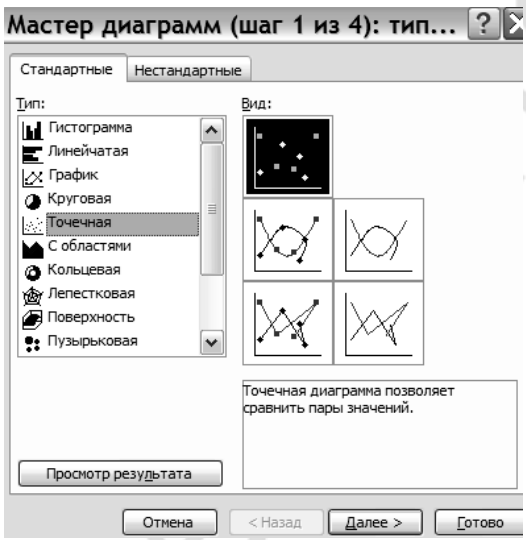


Рис. 109

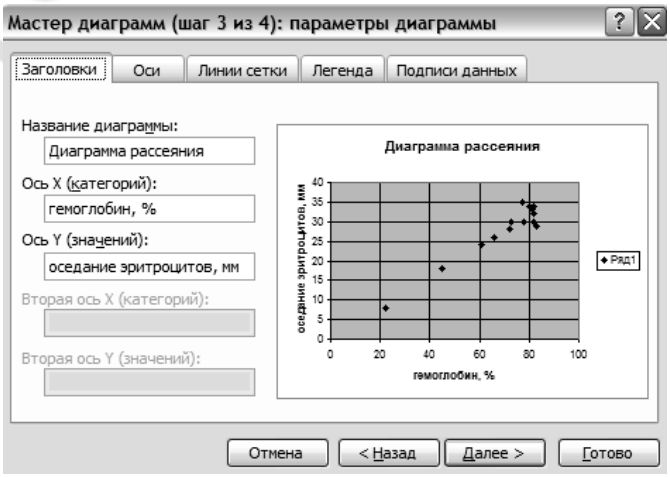


Рис. 110

– на вкладке **Линии сетки** установите флажки *основные линии* в разделах *Ось X и Ось Y*;

– на вкладке **Легенда** снимите флажок **Добавить легенду**, нажмите кнопку **Далее>**;

– в появившемся окне выбора места расположения диаграммы включите переключатель **имеющемся** и нажмите кнопку **Готово**.

2. Снимите заливку области построения диаграммы, для чего:

– выделите область построения (сетку) щелчком правой кнопки мыши;

– в контекстном меню выберите **Формат области построения**;

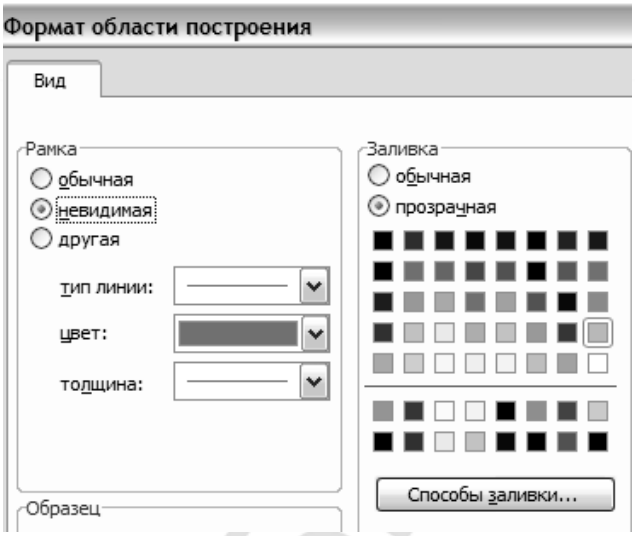


Рис. 111

– на вкладке **Вид** (рис. 111) в области *Заливка* над палитрой установите переключатель прозрачная;

– в области *Рамка* установите переключатель невидимая нажмите **ОК**.

3. Добавьте линию и уравнение регрессии, для чего:

– наведите указатель мыши на точки диаграммы, выделите их как показано на рис. 112, нажмите правую кнопку мыши для вызова контекстного меню;

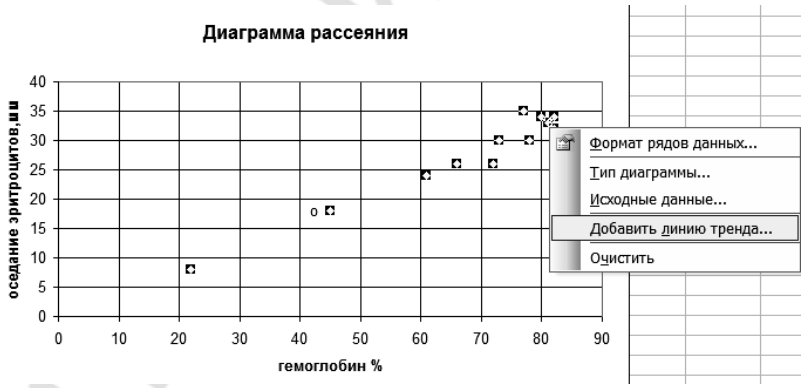


Рис. 112

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

– в окне **Линия тренда** на вкладке *Тип* выберите *Линейная*, на вкладке Параметры (рис. 113) поставьте флажки *Показывать уравнение* и *Поместить величину достоверности аппроксимации*[1].

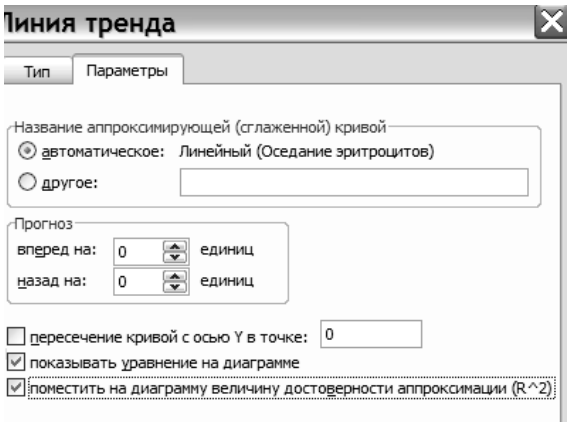


Рис. 113

Результат ваших действий показан на рис. 114.

Так как корреляционное поле вытянуто, то корреляционная связь

[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.

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 *У* прямая, т. к. *r* = 0,96 связь сильная.

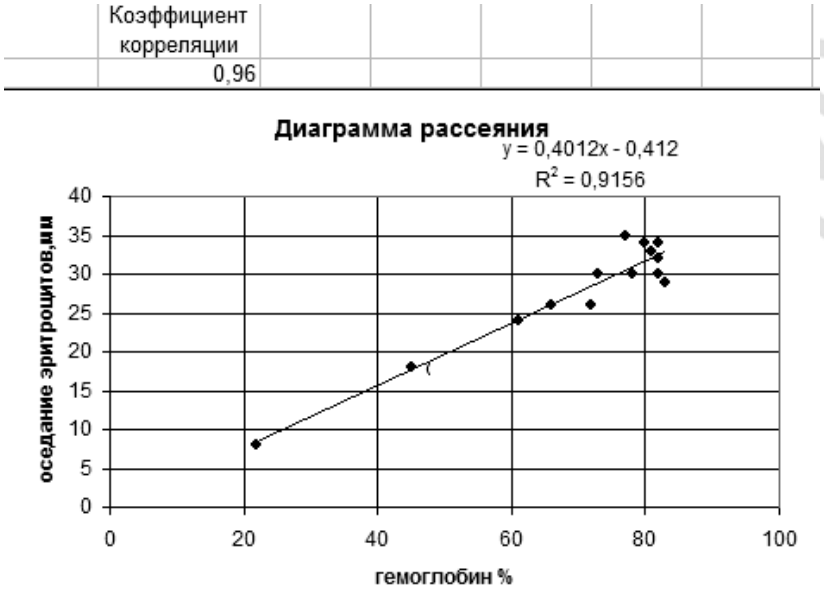


Рис. 114

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

Литература

1. *Афифи,* А. Статистический анализ. Подход с использованием ЭВМ / А. Афифи, С. Эйзен. М. : Мир, 1982. 488 с.

2. *Боровиков, В.* Statistica. Искусство анализа данных на компьютере : для профессионалов / В. Боровиков. СПб. : Питер, 2001. 656 с.

3. *Вентцель, Е. С.* Теория вероятностей / Е. С. Вентцель. М. : Наука, 1969. 576 с.

4. *Гмурман, В. Е.* Руководство к решению задач по теории вероятностей и математической статистике / В. Е. Гмурман. М. : Высшая школа, 2001. 400 с.

5. *Гмурман, В. Е.* Теория вероятностей и математическая статистика / В. Е. Гмурман. М. : 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. Н. 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. М. : Медицина, 2006. 528 с.

9. *Медик, В. А.* Статистика в медицине и биологии : рук. в 2 т. Т. 1. Теоретическая статистика / В. А. Медик, М. С. Токмачев, Б. Б. Фишман. М. : 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. М. : Инфра-М, 1998. 528 с.

12. *Фигурин, В. А.* Теория вероятностей и математическая статистика / В. А. Фигурин, В. В. Оболонкин. Минск : Новое знание, 2000. 206 с.

13. *Юнкеров, В. И.* Математико-статистическая обработка данных медицинских исследований / В. И. Юнкеров, С. Г. Григорьев. СПб. : ВМедА, 2002. 266 с.

Оглавление