МИНИСТЕРСТВО СПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный университет физической культуры,
спорта, молодежи и туризма (ГЦОЛИФК)»

Иркутский филиал РГУФКСМиТ

Компьютерная обработка данных

Сборник задач, упражнений и тестовых заданий

Иркутск, 2020

Рецензент канд. физ.-мат. наук, доцент О.В. Леонова

Рецензент канд. мед. наук, доцент Н.П. Гаськова

Компьютерная обработка данных. Сборник задач, упражнений и тестовых заданий. Учебное пособие/ Н.В. Мамонова. – Иркутск: Изд-во ___, 2020 – 80с.

Содержатся теоретические сведения, задачи, упражнения, контрольные и тестовые задания по компьютерной обработке данных. Излагаемый материал охватывает основные разделы программы курса «Компьютерная обработка данных экспериментального исследования». Кроме необходимых теоретических сведений приводятся примеры решения стандартных задач.

Рекомендуется для студентов очного и заочного отделения Иркутского филиала РГУФК СМиТ

Оглавление

Предисловие

Компьютерная обработка данных основана на знании математической статистики – это является важнейшим инструментом современного специалиста.

Большинство современных специальностей (в том числе и гуманитарных) требует знания методов математической статистики и умений обрабатывать статистические данные. Для студента вуза физической культуры трудность освоения этого раздела заключается в следующем противоречии. С одной стороны, изучение теории вероятностей и математической статистики требует времени, которого в программах вузов этого профиля явно недостаточно. С другой – работа с любым пакетом, производящим статистическую обработку, невозможна без знания математической терминологии и математической постановки задач. Наконец, без математического языка невозможно изучение специальной литературы и опубликование результатов.

Математическая статистика – это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Компьютерная обработка исследуемых данных является логичным следствием применения пакетов прикладных программ.

В области физической культуры и спорта такие методы также применимы. Предположим, что тренер или исследователь предлагает новую методику подготовки спортсменов и ему необходимо оценить её эффективность. Для этой цели он выполняет следующие действия. Набираются две группы испытуемых: контрольная и экспериментальная, которые примерно одинаковы по признакам (факторам), имеющих какое-либо значение для цели исследования (пол, возраст, спортивная квалификация и т. п.). Контрольная группа тренируется по традиционной методике, а экспериментальная – с применением предлагаемых нововведений. После проведения тренировок (а может быть и в процессе тренировок) проводятся обследования испытуемых контрольной и экспериментальной групп, результаты которых изучаются и сравниваются.

Итак, тренер (или исследователь) получает результаты измерений в виде большого объема числовых данных. Для определения эффективности новой методики тренировок, ему необходимо решить множество задач: обработать – представить для исследования – полученный статистический материал, измерить степени влияния разных признаков (факторов) на эффективность тренировок, извлечь из него полезную информацию, и, наконец, сделать выводы. Все эти задачи решаются методами математической статистики. Следует отметить, что без них сегодня не обходится практически ни одна наука, входящая в образование специалиста физической культуры: анатомия, биохимия, физиология, биомеханика, психология, спортивная медицина, педагогика спорта и многие другие.

Современные информационные технологии, основанные на использовании ПК, позволяют быстро решать необходимые математические задачи. Однако, во-первых, задачи обработки данных необходимо сформулировать; во-вторых, ввести их в компьютер на соответствующем «математическом» языке; в-третьих, провести вычисления, и, в-четвертых, интерпретировать полученные результаты, сделав из них правильные выводы.

Любой специалист в области физической культуры и спорта повышает свою квалификацию через изучение литературы по специальности. В большинстве публикаций представляются не исходные данные, а результаты статистической обработки. Поэтому знание терминологии, задач и алгоритмов математической статистики сегодня необходимо.

Задачи и упражнения настоящего сборника затрагивают практические аспекты как основ и задач традиционной математической статистики: теории оценивания и проверки гипотез, так и ее прикладных разделов: дисперсионного, корреляционного и регрессионного анализов. Каждый раздел содержит краткие теоретические сведения и иллюстративные примеры. Задания для самостоятельной работы могут выполнятся как «вручную», так и на компьютере с использованием статистических пакетов (Статграф, Excel/Статистика).

В конце пособия приведены таблицы распределений, список рекомендуемой литературы, а также задания для домашних расчетно-графических работ.

Глава I. Элементы описательной статистики

1.1. Эмпирические распределения и их графические представления

Выборка $\{x_1, x_2, \ldots, x_n\}$объема n имеющихся в нашем распоряжении значений исследуемой случайной величины X (случайного признака) является той исходной информацией, на основании которой строятся выводы о свойствах изучаемой генеральной совокупности в целом и, в частности, составляется представление о функции и ряде распределения или плотности анализируемого закона распределения вероятностей.

Упорядоченная по величине последовательность выборочных значений $x_1^{(n)} \le x_2^{(n)} \le \ldots \le x_n^{(n)}$ называется вариационным рядом. Среди членов вариационного ряда могут быть совпадающие между собой значения. Если через $n_1, n_2, \ldots, n_r$ обозначить число повторений всех несовпадающих значений выборки, то получим два ряда чисел:

x_i	x_1	x_2	$\ldots$	x_r
n_i	n_1	n_2	$\ldots$	n_r

$$, \quad \sum_{i=1}^{r} n_i = n. \qquad (1)$$

Первый ряд содержит различные выборочные значения, расположенный в порядке возрастания. Числа второго ряда показывают количество повторений каждого из этих значений в выборке и называются частотами. Ряд (1) называется точечным вариационным рядом, что соответствует дискретной вариации признака, или эмпирическим распределением признака по частотам.

От распределения частот (т. е. ряда (1)) можно перейти к распределению относительных частот $w_i = \dfrac{n_i}{n}, \sum\limits_{i=1}^{r} w_i = 1,$ заданных в виде доли w_i или в виде процента $w_i \cdot 100\%$, $\sum\limits_{i=1}^{r} w_i \cdot 100\% = 100\%$:

x_i	x_1	x_2	$\ldots$	x_r
w_i	w_1	w_2	$\ldots$	w_r

$$, \quad \sum_{i=1}^{r} w_i = 1. \qquad (2)$$

Вариационный ряд (2), построенный по относительным частотам, является статистической аппроксимацией теоретического ряда распределения вероятностей случайной величины X.

Если объем выборки n велик ($n \ge 50$) и/или число различных значений признака велико, (или мы имеем дело с непрерывной случайной величиной, или дискретной, число возможный значений которой достаточно велико), то часто удобнее, с точки зрения дальнейшей статистической обработки результатов наблюдений, перейти к интервальному вариационному ряду или группированной выборке. Этот переход осуществляется следующим образом:

1) отмечаются наименьше x_{min} и наибольшее x_{max} значения в выборке;

2) весь диапазон $[x_{min}; x_{max}]$ разбивается на k равных интервалов группирования (количество интервалов k не должно быть меньше 8-10 и больше 20-25), выбор числа интервалов существенно зависит от объема выборки n; для примерной ориентации в выборе k можно воспользоваться приближенной формулой

$$k \approx 1+log_2 n, \text{ либо } k \approx 1+1{,}45\, ln\, n;$$

3) определяется величина шага или ширина интервала группирования h, для чего вариационный размах $R = x_{max} - x_{min}$ делится на число интервалов k:

$$h = \frac{x_{\max} - x_{\min}}{k};$$

4) находятся крайние точки каждого из интервалов: $C_0 = x_{\min}$, $C_1 = C_0 + h,\ C_2 = C_1 + h, \ldots, C_k = C_{k-1} + h,$ а также их середины: $x_1^*, x_2^*, \ldots, x_k^*$;

5) подсчитываются числа выборочных данных, попавших в каждый из интервалов: $n_1, n_2, \ldots, n_k$ (очевидно, $n_1 + n_2 + \ldots + n_k = n$); выборочные данные, попавшие на границы интервалов, либо равномерно распределяются по двум соседним интервалам, либо относятся только к какому-либо одному из них, например, к левому.

Таким образом, следуя этой методике, от ряда (1) или (2) можно перейти к интервальному вариационному ряду

$C_{i-1} - C_i$	$C_0 - C_1$	$C_1 - C_2$	$\ldots$	$C_{k-1} - C_k$
n_i	n_1	n_2	$\ldots$	n_k

$$\sum_{i=1}^{k} n_i = n. \qquad (3)$$

От интервального ряда (3) можно вновь перейти к точечному, если в качестве значения случайной величины, соответствующего i - му интервалу, взять его середину $x_i^* = (C_{i-1} + C_i)/2$. Получим ряд

x_i^*	x_1^*	x_2^*	$\ldots$	x_k^*
n_i	n_1	n_2	$\ldots$	n_k

$$\sum_{i=1}^{k} n_i = n. \qquad (4)$$

В некоторых задачах от ряда (1) или (4) целесообразно перейти к ряду, содержащему кумулятивные или накопленные частоты m_i. Накопленная (интегральная или кумулятивная) частота m_i значения x_i получается суммированием частот значений, предшествующих данному, с частотой n_i, т.е. $m_i = n_1 + n_2 + \ldots + n_i$. Накопленная частота крайнего правого значения (или максимального элемента выборки) равна объему выборки n. Несмотря на видимую несхожесть, ряды (1) – (4) отражают одно и то же фактическое распределение признака.

Замечание. Предложенную процедуру построения вариационных рядов ни в коем случае не следует считать единственно возможной. Количество интервалов, их длины, а также расположение интервалов относительно выборочного материала могут варьироваться по усмотрению исследователя в зависимости от решаемых задач.

Для наглядного представления вариационные ряды изображают в виде графиков. Наиболее распространенными способами представления эмпирических данных является гистограмма, полигон частот или относительных частот и полигон накопленных частот или кумулятивная кривая – кумулята.

Гистограмма состоит из последовательности примыкающих друг к другу прямоугольников (рис.1). Ширина этих прямоугольников равна ширине интервалов группирования h и откладывается по оси абсцисс, а высота откладывается по оси ординат и пропорциональна частоте n_i или относительной частоте w_i. В первом случае имеем гистограмму частот с высотами прямоугольников, равными n_i / h, и общей площадью, равной объему выборки n. Во втором – гистограмму относительных частот с высотами прямоугольников $n_i / n \cdot h$, и общей площадью, равной 1. Ступенчатая ломаная, ограничивающая в этом случае сверху построенную фигуру, является статистической аппроксимацией кривой распределения или графика теоретической функции плотности вероятности *f(x)* исследуемой случайной величины *X*. Эту же аппроксимацию мы получим, если через середины верхних оснований прямоугольников проведем плавную линию (пунктир).

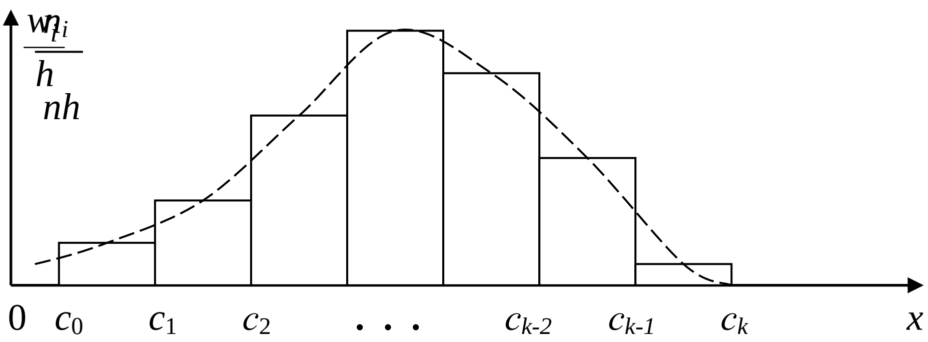


Рис. 1

Полигон частот или относительных частот представляет собой многоугольник с вершинами в точках (x_i, n_i) или (x_i, w_i) (рис.2).

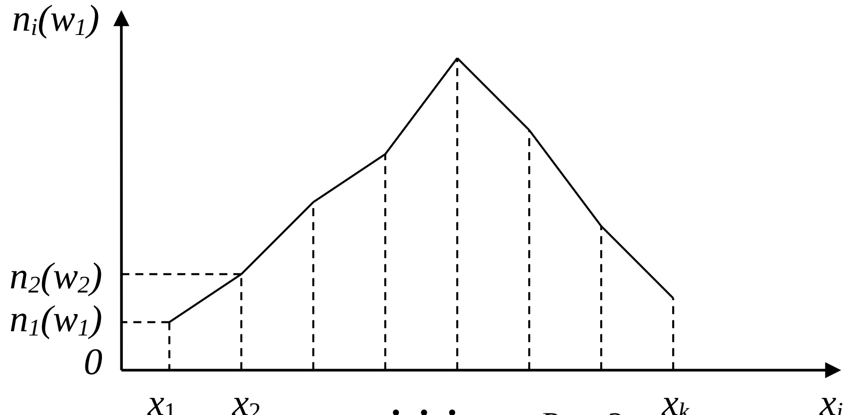


Рис. 2

При изображении полигона частот или относительных частот интервального вариационного ряда вершины многоугольника расположены в точках с абсциссами, соответствующими срединным значениям интервалов x_i^*, и ординатами, равными частоте n_i или относительной частоте w_i.

Полигон накопленных частот (кумулята) получается изображением в прямоугольной системе координат вариационного ряда с накопленными частотами. При построении кумуляты дискретного признака на ось абсцисс наносятся значения признака – элементы выборки x_i. Ординатами служат вертикальные отрезки – накопленные частоты m_i (рис.3).

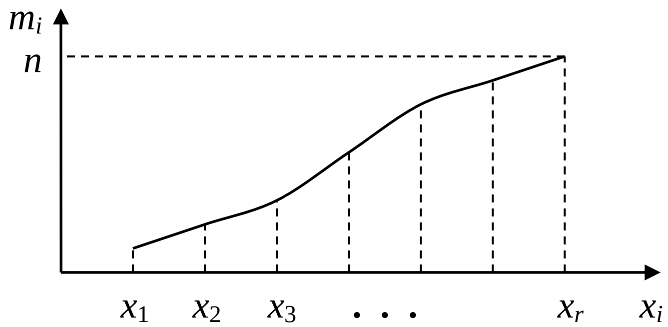


Рис.3

В теории вероятностей одной из форм задания закона распределения случайной величины X является функция распределения $F(x)$. В математической статистике ее эмпирическим аналогом будет статистическая или эмпирическая функция распределения $F^*(x)$.

Для выборки, представленной рядом (1), эмпирическая функция распределения $F^*(x)$ запишется как

$$F^*(x) = \begin{cases} 0, & x \le x_1, \\ \dfrac{n_1}{n}, & x_1 < x \le x_2, \\ \dfrac{n_1 + n_2}{n}, & x_2 < x \le x_3, . \\ \dots & \dots \quad \dots \\ 1, & x > x_r. \end{cases}$$

График эмпирической функции распределения представляет собой ступенчатую линию со скачками в точках, определяемых элементами выборки (рис.4).

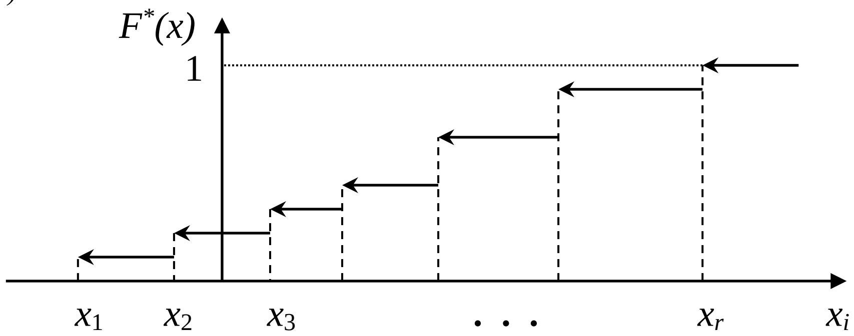


Рис. 4

Проиллюстрируем построение вариационных рядов, их графиков, а также эмпирической функции распределения на следующем примере.

<u>Пример 1</u>. Изучается спринтерский бег X (с). Для этого обследовано n = 100 детей в возрасте 15-16 лет. Результаты наблюдений представлены ниже:

13,9	15,3	15,0	15,4	14,8	14,9	16,0	14,7	15,8	16,4
15,4	13,6	15,4	13,5	15,2	15,0	16,6	15,8	13,8	15,1
16,6	16,9	14,7	13,6	16,0	16,0	16,3	14,1	14,8	13,9
14,7	15,2	15,3	17,1	14,1	14,3	15,6	16,4	16,1	15,9
15,6	15,4	14,9	14,6	15,6	15,2	13,0	13,7	14,2	15,2
16,6	13,6	14,4	15,7	14,8	15,3	15,1	15,0	15,8	16,8
13,6	15,0	15,0	12,5	13,9	13,9	16,4	15,4	15,0	15,1
14,8	15,9	14,5	14,4	15,4	13,8	15,4	15,8	13,4	14,8
15,5	17,3	14,5	13,2	15,4	16,0	14,8	13,8	14,1	15,8
13,8	14,2	14,6	15,6	14,6	16,7	15,5	14,4	14,7	14,1

1. По данным выборки построить точечный вариационный ряд, распределив значения x_i по частотам n_i (ряд 1).
2. От ряда 1 перейти к интервальному ряду (ряд 2).

3. От ряда 2 перейти к точечному ряду, распределив значения по частотам (ряд 3) и относительным частотам в виде доли и в виде процента (ряд 4).

4. Построить: а) гистограмму относительных частот для ряда 2; б) полигон частот для ряда 3; в) кумулятивную кривую для ряда 3.

5. Найти эмпирическую функцию распределения случайной величины X, используя ряд 3, и построить ее график.

Решение.

1. Для того чтобы построить точечный вариационный ряд, необходимо расположить наблюдаемые значения x_i в порядке их возрастания и относительно каждого x_i указать частоту n_i, т.е. количество повторений x_i в выборке, при этом сумма всех частот равна объему выборки n.

Ряд 1:

x_i n_i	12,5 1	13,0 1	13,2 1	13,4 1	13,5 2	13,6 3	13,7 1	13,8 4	13,9 4	14,1 4
x_i n_i	14,2 2	14,3 1	14,4 3	14,5 2	14,6 4	14,7 4	14,8 5	14,9 2	15,0 6	15,1 3
x_i n_i	15,2 4	15,3 4	15,4 7	15,5 2	15,6 4	15,7 1	15,8 5	15,9 2	16,0 4	16,1 1
x_i n_i	16,3 1	16,4 3	16,6 3	16,7 1	16,8 1	16,9 1	17,1 1	17,3 1		

Здесь объем выборки $n = \sum n_i = 100$, а число различных значений $r = 38$.

2. Так как объем выборки велик и число различных значений исследуемого случайного признака также велико, то целесообразно перейти от точечного ряда 1 к интервальному. Такой переход осуществляется по изложенной выше методике следующим образом:

а) отмечаются наименьшее $x_{\min} = 12{,}5$ и наибольшее $x_{\max} = 17{,}3$ значения в выборке;

б) весь обследованный диапазон [12,5;17,3] разбивается на число интервалов k, где $k \approx 1 + \log_2 n$. В нашем примере n =100 и k=8;

в) определяется величина шага или ширина интервала группирования h:

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{17{,}3 - 12{,}5}{8} = \frac{4{,}8}{8} = 0{,}6;$$

г) отмечаются крайние точки каждого из интервалов C_{i-1}, C_i в порядке возрастания, а также подсчитываются числа выборочных данных, попавших в каждый из интервалов $n_1, n_2, \ldots, n_k$ (очевидно, здесь $n_1 + n_2 + \ldots + n_8 = 100$):

Ряд 2:

$C_{i-1}-C_i$	12,5 – 13,1	13,1 – 13,7	13,7 – 14,3	14,3 – 14,9
n_i	2	8	15	20
$C_{i-1}-C_i$	14,9 – 15,5	15,5 – 16,1	16,1 – 16,7	16,7 – 17,3
n_i	26	17	8	4

3. Для того чтобы от интервального ряда 2 перейти вновь к точечному, необходимо отметить середины интервалов x_i^* и сопоставить им частоты n_i или относительные частоты w_i. Распределение производительности труда по частотам запишется в виде ряда 3, а распределение по относительным частотам в виде ряда 4:

Ряд 3:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0
n_i	2	8	15	20	26	17	8	4

, $\sum n_i = 100$

Ряд 4:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0
w_i	0,02	0,08	0,15	0,20	0,26	0,17	0,08	0,04
$w_i \cdot 100$	2	8	15	20	26	17	8	4

$\sum w_i = 1$, $\sum w_i 100\% = 100\%$

4. Гистограмма относительных частот для ряда 2 изображена на рис. 5.

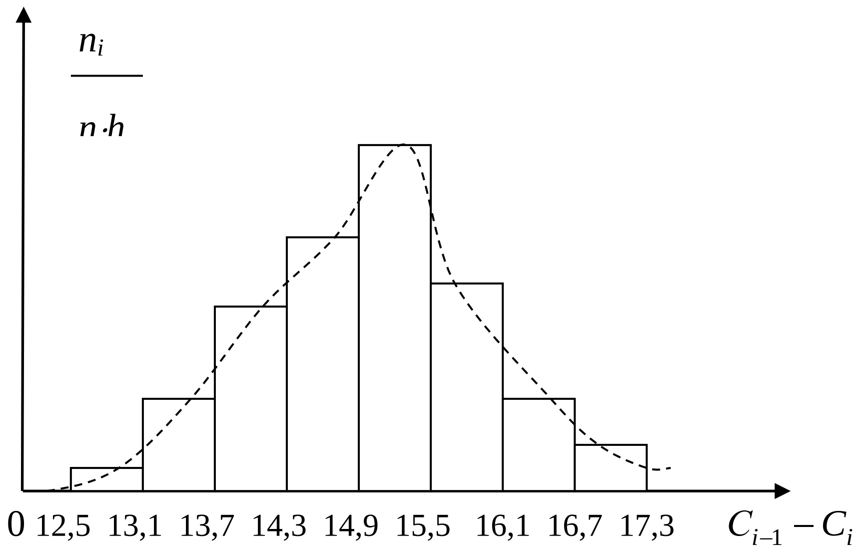


Рис. 5

Полигон частот показан на рис.6.

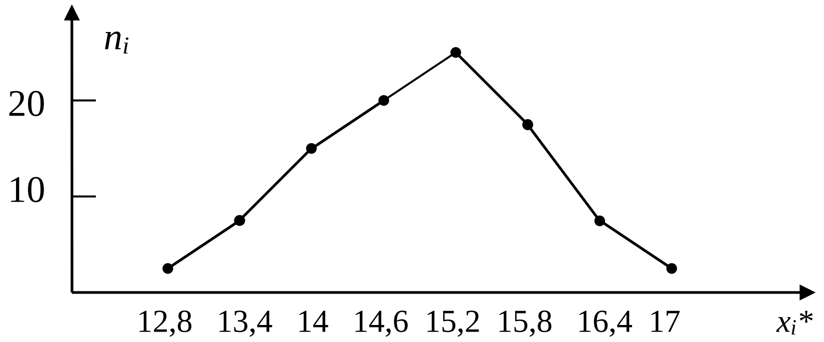


Рис. 6

Для построения кумуляты представим ряд 3 по накопленным частотам m_i:

x_i^*	12,8	13,4	14,0	14,6	15,2	15,8	16,4	17,0
m_i	2	10	25	45	71	88	96	100

Тогда кумулятой будет плавная кривая, изображенная на рис.7.

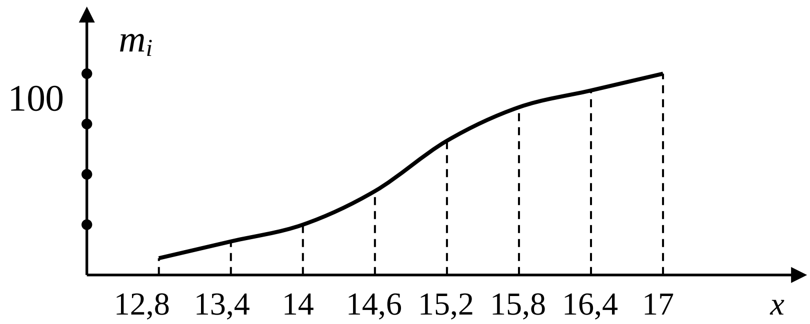


Рис. 7

5. Эмпирическая функция распределения для ряда 3 запишется так:

$$F^*(x) = \begin{cases} 0, & x \le 12{,}8, \\ 0{,}02, & 12{,}8 < x \le 13{,}4, \\ 0{,}10, & 13{,}4 < x \le 14{,}0, \\ 0{,}25, & 14{,}0 < x \le 14{,}6, \\ 0{,}45, & 14{,}6 < x \le 15{,}2, \\ 0{,}71, & 15{,}2 < x \le 15{,}8, \\ 0{,}88, & 15{,}8 < x \le 16{,}4, \\ 0{,}96, & 16{,}4 < x \le 17{,}0, \\ 1, & x > 17{,}0. \end{cases}$$

Здесь, например, значение функции $F^*(x)$, равное 0,02, найдено как $\frac{2}{100}$, так как значение $X < 13{,}4$, а именно $x_1 = 12{,}8$ наблюдалось 2 раза; значения $X < 14{,}0$, а именно $x_1 = 12{,}8$ и $x_2 = 13{,}4$ наблюдались 2 + 8 = 10 раз, следовательно, $F^*(x)$=10/100 =0,10 при $13{,}4 < x \le 14{,}0$ и т. д.

График $F^*(x)$ изображен на рис. 8.

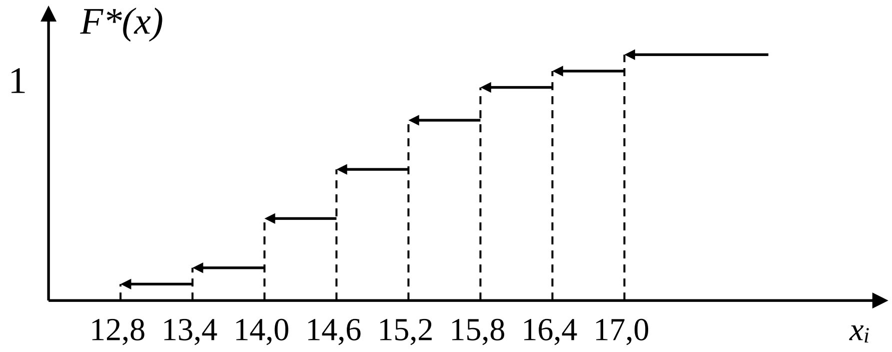


Рис. 8

Задачи

В задачах 1 – 4 построить точечные вариационные ряды, распределив значения исследуемого признака по частотам и относительным частотам в виде доли и в виде процента; изобразить графически полученные эмпирические распределения.

1. Размеры 45 пар мужской обуви, проданных магазином за день:

39	41	40	42	41	40	42	44	40	43	42	41
43	39	42	41	42	39	41	37	43	41	38	43
42	41	40	41	38	44	40	39	41	40	42	40
41	42	40	43	38	39	41	41	42			

2. На вступительных экзаменах 60 абитуриентов получили следующие баллы:

20	19	22	24	21	18	23	17	20	16	15	23
23	21	22	21	24	21	18	23	21	19	20	24
21	20	18	21	19	25	17	22	20	16	22	18
20	17	21	17	19	20	19	20	21	20	21	18
22	23	21	25	22	20	19	21	24	22	23	21

3. Приведены данные о количестве членов семей в 50 обследованных фермерских хозяйствах:

2	5	5	6	3	2	5	6	5	6	6	6	4
3	3	5	7	3	3	5	5	5	4	5	6	4
4	4	4	7	4	4	3	5	3	7	4	6	6
4	7	4	4	6	7	6	3	3	5	8	5	

4. Имеются следующие данные о тарифных разрядах 50 рабочих одного из цехов завода:

3	5	6	3	2	4	3	5	5	6
4	3	2	3	4	5	4	2	4	6
5	3	4	5	4	3	3	6	2	3
4	6	3	4	4	5	4	5	3	4
2	6	3	4	5	3	4	4	5	4

В задачах 5 – 8 построить интервальные вариационные ряды, изобразить их в виде гистограмм. Записать аналитически и построить график эмпирической функции распределения.

5. Даны результаты 50 наблюдений за жирностью молока (%):

3,86	4,06	3,67	3,97	3,76	3,61	3,96	4,04	3,84	3,94
3,98	3,57	3,87	4,07	3,99	3,69	3,76	3,71	3,94	3,82
4,16	3,76	4,00	3,46	4,08	3,88	4,01	3,93	3,71	3,81
4,02	4,17	3,72	4,09	3,78	4,02	3,73	3,52	3,85	3,92
4,18	4,26	4,03	4,14	3,72	4,33	3,82	4,03	3,62	3,91

6. Имеются следующие данные об урожайности озимой пшеницы в 40 обследованных хозяйствах:

27,1	18,2	16,3	22	24,3	24,8	33,0	27,3	28,5	15,1
19,5	28,1	25,1	26,7	28,4	29,6	23,7	18,0	31,0	19,8
26,0	23,5	20,2	25,1	22,8	27,0	20,4	24,0	29,5	22,9
19,9	27,0	25,3	23,9	21,5	23,1	21,1	22,6	25,8	23,8

7. Имеются следующие данные о стоимости основных фондов у 50 предприятий (млн. руб.):

9,4	8,0	6,3	10,0	15,0	8,2	7,3	9,2	5,8	8,7
5,2	13,2	8,1	7,5	11,8	14,6	8,5	7,8	10,5	6,0
5,1	6,8	8,3	7,7	7,9	9,0	10,1	8,0	12,0	14,0
8,2	9,8	13,5	12,4	5,5	7,9	9,2	10,8	12,1	12,4
12,9	12,6	6,7	9,7	8,3	10,8	15,0	7,0	13,0	9,5

8. Дана выборка затрат (в копейках) на рубль продукции (работ, услуг) по 100 предприятиям:

61,55	61,59	62,09	63,08	63,97	64,74	65,07
67,12	68,10	69,38	70,21	70,21	70,36	71,25
71,86	72,00	72,39	72,41	72,46	72,50	72,80
72,84	73,44	74,93	75,46	75,65	77,13	77,37
77,64	77,86	77,93	78,03	78,28	78,74	78,97
79,07	79,10	79,34	79,34	79,34	79,40	79,49
79,70	80,02	80,26	80,56	80,65	80,69	81,13
81,32	81,40	81,54	81,85	82,27	82,71	82,74
82,78	83,03	83,05	83,59	83,68	83,74	83,78
83,96	84,98	85,18	85,32	85,64	85,71	85,84
86,01	86,03	86,05	86,11	86,48	86,94	86,98
87,38	87,47	87,59	87,89	88,03	88,04	88,11
88,24	88,89	90,34	90,40	90,58	90,73	90,76
92,51	92,72	92,94	94,58	95,06	95,73	96,11
96,34	96,55					

1.2. Числовые характеристики эмпирических распределений

Исчерпывающие сведения об интересующем нас законе распределения вероятностей дают вариационные ряды, их графические представления, а также эмпирическая функция распределения. Однако нередко при практическом изучении генеральной совокупности этого бывает недостаточно, и требуется охарактеризовать имеющуюся совокупность значений некоторыми количественными показателями. К таким показателям или числовым характеристикам выборки относятся меры положения, меры рассеяния и меры формы.

Характеристики или меры положения. Существует несколько характеристик, применяемых для описания характера расположения распределений: среднее (арифметическое, геометрическое и гармоническое), медиана, мода, а также выборочные квантили.

Арифметическое (или выборочное) среднее $\bar{x}$ (или $\bar{x}_в$) для несгруппированной выборки $\{x_1, x_2, \ldots, x_n\}$ объема n определяется формулой

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,. \tag{5}$$

В случае выборки, представляемой рядом вида (1), выборочное среднее равно:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{r} n_i x_i \,. \tag{6}$$

Выборочное среднее представляет собой значение, относительно которого может быть «сбалансировано» все эмпирическое распределение (фактически, это абсцисса центра масс гистограммы). Эта характеристика является одной из наиболее употребительных статистических мер: многие средние показатели в экономике подсчитываются по формулам (5), (6).

<u>Среднее геометрическое</u> $\overline{x}_{геом}$ определяется как

$$\overline{x}_{геом} = \sqrt[n]{x_1 x_2 \dots x_n} \, ,$$

либо

$$\overline{x}_{геом} = \sqrt[n]{x_1^{n_1} x_2^{n_2} \dots x_r^{n_r}} \, .$$

Геометрическое среднее находит применение при оценке темпов изменения величин, например, при расчетах индексов цен. Геометрическое среднее следует применять прежде всего тогда, когда среднее значение должно быть рассчитано из значений, заданных через некоторые равные промежутки времени; $\overline{x}_{геом}$ применяется, когда переменная меняется во времени с приблизительно постоянным соотношением между измерениями. К этому случаю относятся многообразные явления роста. Прирост населения во времени, изменение числа пациентов или эксплуатационные расходы – вот известные примеры подобного типа явлений.

<u>Среднее гармоническое</u> $\overline{x}_{гарм}$ задается соотношением

$$\overline{x}_{гарм} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

или

$$\overline{x}_{гарм} = \frac{1}{\frac{1}{n}\sum_{i=1}^{r}\frac{n_i}{x_i}} .$$

Область применения гармонического среднего весьма ограничена. В экономике, в частности, пользуются иногда гармоническим средним при анализе средних норм времени, а также в некоторых видах индексных расчетов, когда суммируемый признак выражен обратной величиной данного признака, т.е. $1/x_1, \quad 1/x_2, \dots, 1/x_n$.

Между тремя средними значениями существует следующее соотношение:

$$\overline{x}_{гарм} \le \overline{x}_{геом} \le \overline{x} \, ,$$

причем равенство справедливо при одинаковых выборочных значениях.

<u>Медиана</u> x_{med} исследуемого признака определяется как его средневероятное значение, т.е. такое значение, для которого

$$P\{X < x_{med}\} = P\{X > x_{med}\} = \frac{1}{2}.$$

При определении выборочного (приближенного) значения медианы имеющиеся в нашем распоряжении наблюдения $x_1, x_2, ..., x_n$ располагают в вариационный ряд и определяют в качестве x_{med} средний (т. е. $\frac{1}{2}(n+1)$-й) член этого ряда, если n нечетно, и любое значение между средними, т.е. $\frac{1}{2}n$–м и $(\frac{1}{2}n+1)$ –м членами этого ряда, если n четно.

При исчислении медианы интервального вариационного ряда вначале находят интервал, содержащий медиану. Медианному интервалу соответствует первая из накопленных частот, превышающая половину объема выборки. Для нахождения медианы при постоянстве плотности внутри интервала, содержащего медиану, используют следующую формулу:

$$x_{med} = x_{med(\min)} + h\frac{n/2 - m_{med-1}}{n_{med}}, \qquad (7)$$

где $x_{med(\min)}$ - нижняя граница медианного интервала; h – интервальная разность; m_{med-1} - накопленная частота интервала, предшествующего медианному; n_{med} - частота медианного интервала.

Медиана может быть определена и графически по кумуляте. Для этого последнюю ординату, равную сумме всех частот, т.е. объему выборки n, делят пополам. Из полученной точки восстанавливают перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения и дает значение медианы (рис. 9).

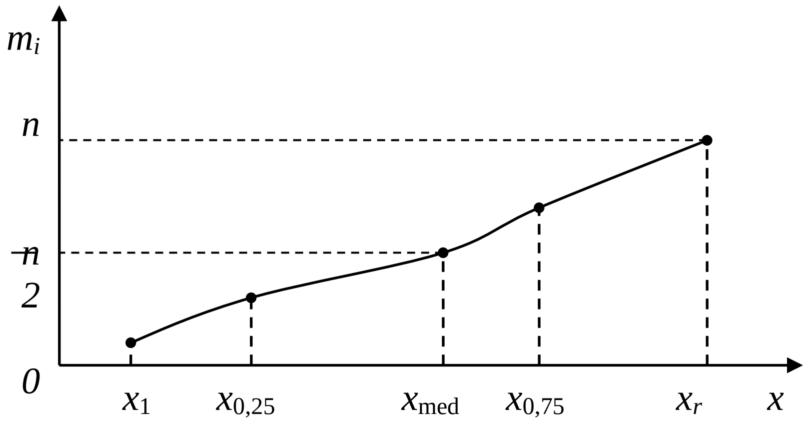


Рис. 9

<u>Модальное значение</u> (или просто *мода*) x_{mod} есть такое значение исследуемого признака, которое чаще всего встречается в данном вариационном ряду. Для дискретного ряда мода определяется по частотам и соответствует выборочному значению с наибольшей частотой. В случае интервального распределения с равными интервалами модальный интервал, т. е. содержащий

моду, определяется по наибольшей частоте, а при неравных интервалах – по наибольшей плотности. Вычисление моды производится по формуле:

$$x_{\text{mod}} = x_{\text{mod(min)}} + h\frac{n_{\text{mod}} - n_{\text{mod}-1}}{2n_{\text{mod}} - n_{\text{mod}-1} - n_{\text{mod}+1}}, \tag{8}$$

где $x_{\text{mod(min)}}$ - нижняя граница модального интервала; h – интервальная разность; n_{mod} - частота модального интервала; $n_{\text{mod}-1}$ - частота интервала, предшествующего модальному; $n_{\text{mod}+1}$ - частота интервала, следующего за модальным.

В случае симметричной плотности среднее значение $\bar{x}$, мода x_{mod} и медиана x_{med} совпадают между собой.

Выборочной квантилью порядка или уровня p называется абсцисса x_p точки, лежащей на кумулятивной кривой и имеющей ординату p (см. рис. 9). Порядок квантили p определяет долю общего числа наблюдений в выборке, результаты которых не превосходят x_p. Квантили порядка 0,25 и 0,75 называют соответственно нижним и верхним квартилями, медиана есть квантиль порядка 0,5, т. е. $x_{0,5} = x_{med}$. Нахождение квартилей осуществляется точно так же, как и определение медианы.

Характеристики или меры рассеяния. Средние величины, характеризующие вариационный ряд одним числом, не учитывают вариацию или разброс значений признака. Для измерения вариации применяется ряд способов.

Вариационный размах *R*, представляющий собой разность между наибольшим и наименьшим значениями в выборке:

$$R = x_{\max} - x_{\min},$$

применяется в качестве приблизительной оценки вариации. Особенно широко используется размах в ряде отраслей промышленности при статистическом изучении качества продукции.

Одной из наиболее часто используемых характеристик рассеяния данных является выборочное среднее квадратическое (стандартное) отклонение:

$$\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

дающее абсолютный разброс значений признака относительно среднего и определяемое таким образом для несгруппированных данных. Если данные сгруппированы, то

$$\sigma_B = \sqrt{\frac{1}{n}\sum_{i=1}^{r}n_i(x_i - \bar{x})^2}.$$

Квадрат этой величины σ^2 называется выборочной дисперсией и обозначается D_B:

$$D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - (\bar{x})^2.$$

Выборочная дисперсия также может использоваться для оценки разброса значений исследуемого признака.

Для оценки относительной изменчивости признака используется коэффициент вариации $V = \frac{\sigma_B}{\bar{x}} \cdot 100\%$, который дает возможность охарактеризовать относительный разброс значений признака вокруг его среднего, выраженный в процентах.

Меры формы. Форма кривой распределения исследуемой случайной величины характеризуется коэффициентами асимметрии и эксцесса, выборочные значения которых определяются формулами

$$A_s = \frac{\mu_3}{\sigma_B^3}, \qquad E_k = \frac{\mu_4}{\sigma_B^4} - 3,$$

где μ_3, μ_4 - центральные эмпирические моменты третьего и четвертого порядков соответственно. Для несгруппированной выборки объема n центральный момент k-го порядка равен: $\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k, \quad k = 1,2,3,4,\ldots.$

Если выборка сгруппирована, то $\mu_k = \frac{1}{n}\sum_{i=1}^{r} n_i (x_i - \bar{x})^k$, $k = 1,2,3,4,\ldots.$ Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю. Поэтому, если для изучаемого распределения эти коэффициенты имеют небольшие значения, то можно предположить близость эмпирического распределения к нормальному закону. Наоборот, большие значения этих характеристик указывают на значительное отклонение от нормального распределения.

Асимметрия служит для характеристики «скошенности» распределения. Если коэффициент асимметрии положительный, то более пологая часть кривой распределения расположена правее моды, если отрицательный – левее (рис. 10).

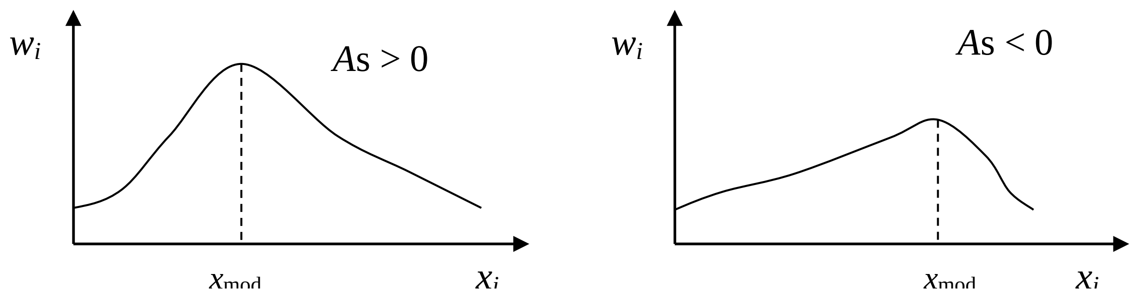


Рис. 10

Для оценки большего или меньшего подъема кривой эмпирического распределения по сравнению с нормальной кривой, используется коэффициент эксцесса. Если $E_k > 0$, то эмпирическая кривая имеет более высокую и «острую» вершину, чем кривая нормального распределения; если $E_k < 0$, то сравниваемая кривая имеет более низкую и плоскую вершину (рис. 11).

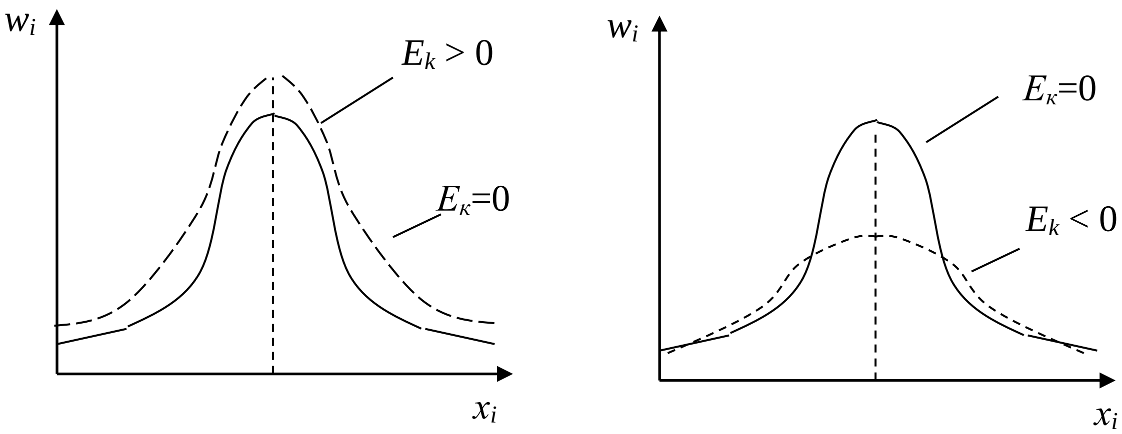


Рис. 11

<u>Пример 2</u>. При изучении временного промежутка по спринтерскому бегу X с. по данным, представленным в примере 1, определить выборочное среднее $\overline{x}$, выборочную дисперсию D_B, выборочное среднее квадратическое отклонение σ_B, коэффициент вариации V, моду x_{mod} и медиану x_{med} по точечному ряду 1 и интервальному ряду 2, а также коэффициенты асимметрии A_s и эксцесса E_k. Проанализировать результаты, полученные в итоге первичной статистической обработки данных, используя решения примеров 1 и 2.

<u>Решение</u>. Для упрощения вычислений расчет характеристик выборки произведем по ряду 3. Для удобства вычислений составим вспомогательную таблицу (табл. 1).

Таблица 1

Вспомогательная таблица для расчета характеристик выборки по сгруппированным данным

k	x_i^*	n_i	$x_i^*\, n_i$	$n_i\,(x_i-\overline{x})$	$n_i\,(x_i-\overline{x})^2$	$n_i\,(x_i-\overline{x})^3$	$n_i\,(x_i-\overline{x})^4$	m_i
1	12,8	2	25,6	- 4,356	9,4874	- 20,6635	45,0051	2
2	13,4	8	107,2	- 12,624	19,9207	- 31,4348	49,6041	10
3	14,0	15	210,0	- 14,670	14,3473	- 14,0316	13,7229	25
4	14,6	20	292,0	- 7,560	2,8577	- 1,0802	0,4083	45
5	15,2	26	396,2	5,772	1,2814	0,2845	0,0632	71
6	15,8	17	268,6	13,974	11,4866	9,4420	7,7613	88
7	16,4	8	131,2	11,376	16,1767	23,0032	32,7106	96
8	17,0	4	68,0	8,088	16,3539	33,0677	66,8628	100

Итог	-	100	1497,8	0,0	91,9117	- 1,4127	216,1383	-

Пользуясь данными табл. 1 и формулой (6), найдем выборочное среднее

$$\overline{x} = \frac{1}{100}\sum_{i=1}^{8} n_i x_i^* = \frac{1497{,}8}{100} = 14{,}978.$$

Для проверки правильности вычисления $\overline{x}$ полезно убедиться в выполнении условия $\sum n_i (x_i^* - \overline{x}) = 0$.

По данным табл. 2 найдем выборочные:

- дисперсию

$$D_B = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \overline{x}) = 0{,}9191;$$

- среднее квадратическое отклонение σ_B =0,9587;
- коэффициент вариации $V = \frac{\sigma_B}{\overline{x}} \cdot 100\% = 6{,}4\%$;
- центральные моменты третьего и четвертого порядков:

$$\mu_3 = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \overline{x})^3 = -0{,}0141;$$

$$\mu_4 = \frac{1}{100}\sum_{i=1}^{8} n_i (x_i^* - \overline{x})^4 = 2{,}1614;$$

- коэффициент асимметрии:

$$A_s = \frac{\mu_3}{\sigma_B^3} = \frac{-0{,}0141}{0{,}8811} = -0{,}0160;$$

- коэффициент эксцесса:

$$E_k = \frac{\mu_4}{\sigma_B^4} - 3 = \frac{2{,}1614}{0{,}8447} - 3 = -0{,}4412.$$

Определим моду и медиану. Мода исследуемой случайной величины X для заданного эмпирического распределения в виде ряда 1 x_{mod} =15,4, так как частота этого значения наибольшая и равна 7. В случае интервального ряда 2 модальному интервалу соответствует наибольшая частота n_{mod}, равная 26. Следовательно,

$x_{\text{mod(min)}}$ =14,9; h =0,6; n_{mod} =26, $n_{\text{mod}-1}$ =20, $n_{\text{mod}+1}$ =17.

$$x_{\text{mod}} = 14{,}9 + 0{,}9\frac{26 - 20}{2 \cdot 26 - 20 - 17} = 15{,}14.$$

Медиану определим как средний член ряда по точеному распределению выборки. В нашем примере n =100, поэтому в качестве медианы берем любое значение между 50-м и 51-м членами ряда 1. Здесь x_{med} =15,0.

Медианному интервалу заданного эмпирического распределения в виде ряда 2 соответствует накопленная частота 71, отсюда $x_{med(\min)}$=14,9; h=0,6; m_{med-1}= 45; n_{med}=26. Используя формулу (7), получим

$$x_{med} = 14{,}9 + 0{,}6\frac{50-45}{26} = 15{,}0154 \approx 15{,}02.$$

Определим медиану графически по кумуляте, представленной на рис. 7. Для этого последнюю ординату, равную объему выборки n=100, поделим пополам. Восстановим перпендикуляр до пересечения с кумулятой. Абсцисса точки пересечения $x_{med}\approx$15 и будет медианой.

Таким образом, среднее время изученной группы детей составила $\bar{x} = 14{,}978$ (с.), абсолютный разброс значений показателя X равен σ = 0,9587 (с.), относительный разброс $V = 6{,}4\%$. Наибольшее время 15,14 (с.), а половина – более 15,02 (с.)

Построенные вариационные ряды 1-3, их графические изображения (рис. 5-8) представляют данные в компактном виде. Кроме этого, имеется возможность получить сведения о законе распределения вероятностей исследуемой случайной величины. Здесь внешний контур гистограммы (рис. 5), графики кумулятивной кривой (рис. 7) и эмпирической функции распределения (рис. 8) свидетельствуют о близости эмпирического распределения к нормальному закону. К этому же выводу можно прийти, сравнивая значения выборочного среднего, моды, медианы. Так как $\bar{x}$, x_{mod} и x_{med} незначительно отличаются друг от друга ($\bar{x}\approx x_{\mathrm{mod}}\approx x_{med}\approx$15,00), есть основание предполагать, что теоретическое распределение симметрично относительно своего среднего значения, что является еще одним доводом в пользу выбора модели нормального закона. И, наконец, близость значений выборочных коэффициентов асимметрии A_s и эксцесса E_k к нулю также свидетельствует в пользу выбора нормального закона распределения для анализируемой случайной величины.

Следовательно, в результате первичной статистической обработки данных мы получили возможность определить некоторые средние показатели интересующего нас признака, а также считать, что случайная величина X – время спринтерского бега детей в возрасте 15-16 лет – распределена по нормальному закону. Нахождение приближенных значений параметров этого закона (оценок) и строгое подтверждение гипотезы о нормальном распределении составляют содержание следующих задач и приемов математической статистики.

Задачи

9. Имеется следующее распределение 60 рабочих по тарифному разряду (X):

x_i	2	3	4	5	6
n_i	8	16	17	12	7

Определить средний тарифный разряд рабочих, моду и медиану.

10. По пяти хозяйствам имеются данные об урожайности зерновых:

№	1	2	3	4	5
x_i	18	20	21	22	25

Рассчитать среднюю гармоническую для всех хозяйств.

11. Автомобиль с грузом от предприятия до склада ехал со скоростью 40 км/ч., а обратно порожняком – со скоростью 60 км/ч. Какова средняя скорость автомобиля?

12. Максимальный размер выигрыша в лотерее составляет 1 000 000 руб. а минимальный – 100 руб. Какова величина среднего выигрыша?

В задачах 13-16 найти моду и медиану, выборочную дисперсию и коэффициент вариации.

13. Обследование качества пряжи на прочность дало следующие результаты:

$C_{i-1} - C_i$	120-140	140-160	160-180	180-200	200-220	220-240	240-260	260-280
n_i	1	6	19	58	53	24	16	3

14. Данные об урожайности ржи на различных участках колхозного поля:

$C_{i-1} - C_i$	9-12	12-15	15-18	18-21	21-24	24-27
n_i	6	12	33	22	19	8

15. Выполнение норм выработки характеризуется следующими данными:

Процент выполнения норм выработки	Число рабочих
90-100	10
100-110	160
110-120	100
120-130	60
130-140	20

16. Дано распределение рабочих по времени, затраченного на обработку одной детали:

Время обработки одной детали, мин.	Число рабочих
2 - 4	42
4 - 6	73
6 - 8	154
8 - 10	205
10 - 12	26

17. Дано распределение семей по числу детей:

Число детей	0	1	2	3	4	5	6
Число семей	10	30	75	45	20	15	6

Найти моду и медиану, выборочные квантили $x_{0,25}$ и $x_{0,75}$. Прокомментировать полученные результаты.

18. Найти моду и медиану, выборочные квантили $x_{0,1}$ и $x_{0,3}$ случайного признака, распределение которого дано в задаче 4.

19. Найти асимметрию и эксцесс следующего распределения признака X:

x_i	2	3	4	5	6	7
n_i	2	6	8	5	3	1

20. Найти асимметрию и эксцесс распределения, заданного в виде ряда:

x_i	- 4	- 3	-2	-1	0	2	3
n_i	1	5	10	7	4	2	1

Контрольные и тестовые задания

1. Вариационный ряд исследуемой случайной величины X, это…
   а) упорядоченная по возрастанию последовательность выборочных значений этой величины;
   б) упорядоченная по убыванию последовательность выборочных значений этой величины;
   в) неупорядоченный набор выборочных значений этой величины;
   г) все значения, которые способна принимать эта величина в случайном порядке.
2. Гистограмма частот строится по…
   а) точечному вариационному ряду;
   б) ряду по относительным частотам;

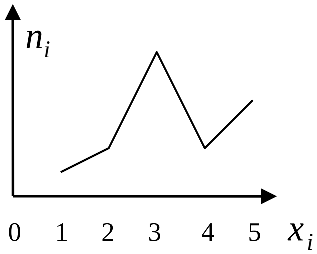
в) ряду по накопленным относительным частотам;
г) интервальному вариационному ряду.

3. Площадь гистограммы относительных частот равна
   а) объему выборки *n*;
   б) единице;
   в) наибольшей частоте;
   г) наибольшей относительной частоте.
4. Выборочное значение является статистическим аналогом…
   а) среднего квадратического отклонения;
   б) дисперсии;
   в) математического ожидания;
   г) коэффициента асимметрии.
5. Коэффициент эксцесса представляет собой характеристику…
   а) остро- или плосковершинности кривой распределения;
   б) односторонней скошенности кривой распределения;
   в) абсолютного разброса исследуемой случайной величины *X*;
   г) относительного разброса.
6. Дана выборка значений случайной величины *X*, представленная в виде ряда:

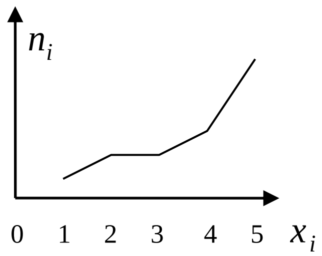
x_i	0	1	2	3
n_i	5	1	20	6

   Определите модальное значение данной величины:
   а) 1; в) 2;
   б) 20; г) 3.
7. Построены полигоны частот различных выборок значений исследуемой случайной величины *X*. Укажите, на каком из графиков модальное значение равно 5?

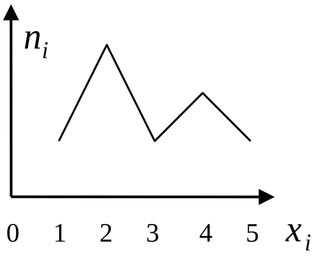
а)



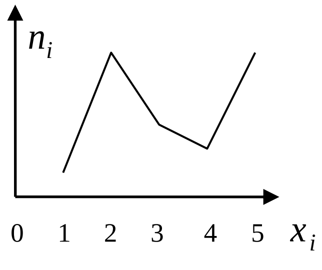
б)



в)



г)



8. Какие характеристики описывают форму кривой распределения?
   а) мода и медиана;
   б) дисперсия и вариационный размах;
   в) асимметрия и эксцесс;
   г) коэффициент корреляции.
9. Наблюдавшиеся значения случайной величины X оказались равными: 3, 0, 4, 3, 6, 0, 3, 1. Чему равно значение эмпирической функции распределения при $x = 5$?
10. Заданы две выборки значений случайной величины
   X: 13, 7, 24, 18, 7, 15;
   Y: 15, 6, 27, 19, 8, 23, 5, 13.
   Для них одинаковой числовой характеристикой является:
   а) квантиль;
   б) медиана;
   в) мода;
   г) начальный момент первого порядка.
11. Распределение предприятий по издержкам обращения (млн. руб.) представлено в виде

$x_i - x_{i+1}$	2 - 6	6 - 10	10 - 14
n_i	3	10	7

   где x_i - объемы издержек обращения,
   n_i - количество предприятий, попавших в i – интервал.
   Чему равна площадь гистограммы частот?
12. Указать неверное утверждение…
   а) медиана – это квантиль уровня 0,5;
   б) эмпирическая функция распределения оценивает теоретическую функцию распределения;
   в) коэффициент асимметрии является характеристикой рассеивания;
   г) математическое ожидание является характеристикой среднего значения случайной величины.
13. Для данных 1, 2, 5, 3, 0, 1, 2, 7, 5, 4, 9, 2, 3, 6, 4 построить статистический ряд по относительным частотам.
14. По выборке получены следующие характеристики: $\bar{x}_B = 60$, $D_B = 400$. Чему равна относительная мера рассеивания вариантов ряда?
15. Определить медиану по данной выборке: 1,5; 8,7; 13; 9; 4; 6; 7,3; 5,9; 11,7; 3,1; 9,8; 12,4.
16. Если на плоскости точки $\left(x_i, \dfrac{n_i}{n}\right)$ соединить отрезками прямых, то получим…
   а) полигон частот;
   б) кумуляту;

в) полигон относительных частот;

г) гистограмму.

17. Сумма отклонений вариант выборки: 8,5; 7,1; 8,7; 6,2; 2,9; 4,4; 6; 5,8; 5,4; 3,5 от среднего арифметического этих чисел равна…

а) 1,2; б) 0; в) 10,8; г) 5,7.

18. Полигон частот представляет собой

а) многоугольник распределения с вершинами $(x_i; n_i)$; $i = 1,.., n$;

б) последовательно примыкающие друг к другу прямоугольники;

в) плавную возрастающую кривую распределения;

г) график точечного вариационного ряда, значения которого распределены по частотам.

19. Укажите правильную формулу расчета выборочного среднего

а) $\frac{1}{n}\sum_{i=1}^{n} n_i$; б) $\frac{1}{n}\sum_{i=1}^{n} x_i$; в) $\frac{1}{n}\sum_{i=1}^{n} m_i$; г) $\frac{1}{n}\sum_{i=1}^{n} x_i n_i$.

Глава 2. Статистическое оценивание параметров

В этом разделе мы изучим методы нахождения точечных и интервальных оценок параметров, характеризующих модели вероятностных распределений. Напомним постановку задачи и свойства точечных оценок.

Пусть мы располагаем исходными статистическими данными – выборкой $\{x_1, \dots, x_n\}$ объема n из исследуемой генеральной совокупности, и пусть изучаемая случайная величина распределена по закону $p(x;\theta)$, где θ – параметр распределения, значение которого неизвестно. Задача оценивания неизвестного параметра θ состоит в построении такой функции $\widehat{\theta} = T(x_1, \dots, x_n)$ от имеющихся в нашем распоряжении данных, которая давала бы в определенном смысле наиболее точное приближенное значение истинного (не известного нам) значения θ. Любую функцию выборки называют статистикой. Статистика $\widehat{\theta}$, используемая в качестве приближенного значения неизвестного параметра θ, называется статистической оценкой. Оценка, полученная в виде одного числа – точки на числовой оси, называется точечной.

Все статистики и статистические оценки являются случайными величинами, принимающими различные значения при переходе от одной выборки к другой (даже в рамках одной и той же генеральной совокупности). Однако значения оценки, подсчитанные по разным выборкам и подверженные случайному разбросу, должны концентрироваться около истинного значения оцениваемого параметра. Это обеспечивается требованиями, предъявляемыми к точечным оценкам, которые формулируются обычно с помощью трех свойств оценок: состоятельности, несмещенности и эффективности.

Состоятельность. Оценка $\widehat{\theta}$ неизвестного параметра θ называется состоятельной, если по мере роста числа наблюдений n (т. е. при $n \to \infty$) она стремится по вероятности к истинному значению θ, т. е. если для $\varepsilon > 0$ $\lim\limits_{n\to\infty} P\{|\widehat{\theta} - \theta| < \varepsilon\} = 1.$

Несмещенность. Оценка $\widehat{\theta} = T(x_1, \dots, x_n)$ неизвестного параметра θ называется несмещенной, если при любом объеме выборки n результат ее усреднения по всем возможным выборкам данного объема приводит к точному истинному значению оцениваемого параметра, т. е. $M\widehat{\theta} = \theta.$

Эффективность. Оценка $\widehat{\theta}$ параметра θ называется эффективной, если среди всех прочих оценок этого параметра она обладает наименьшей мерой случайного разброса относительно его истинного значения. Эффективность является решающим свойством, определяющим качество оценки и состоящим в том, что из всех возможных несмещенных оценок $\widehat{\theta}_i$ параметра θ, вычисленных по выборкам одного и того же объема, выбирается та, которая имеет наименьшую дисперсию, т. е. $\widehat{\theta}$ – эффективная оценка параметра θ, если $D\widehat{\theta} = \min_i D\widehat{\theta}_i$.

На практике при оценке параметров не всегда оказывается возможным одновременное выполнение требований несмещенности, эффективности и состоятельности оценки. Определяют оценки, удовлетворяющие указанным требованиям асимптотически, т.е. при $n \to \infty$.

2.1. Метод максимального правдоподобия

Оценкой максимального правдоподобия (МП-оценкой) называется оценка $\widehat{\theta}_{МП} = T(x_1, x_2, \ldots, x_n)$, которая обращает в максимум функцию правдоподобия

$$L(\theta) = p(x_1, \theta)\, p(x_2, \theta) \cdots p(x_n, \theta) = \prod_{i=1}^{n} p(x_i; \theta). \tag{9}$$

Если функция правдоподобия $L(\theta)$ дифференцируема по θ, то оценку $\widehat{\theta}_{МП}$ находят, решая относительно θ уравнение правдоподобия

$$\frac{\partial L(\theta)}{\partial \theta} = 0,$$

или систему уравнений правдоподобия

$$\frac{\partial L(\theta_1, \ldots, \theta_k)}{\partial \theta_j} = 0, \quad j = 1, \ldots, k,$$

в случае многих неизвестных параметров.

Для удобства можно находить максимум не функции правдоподобия, а логарифмической функции правдоподобия

$$l(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln\, p(x_i; \theta)$$

в силу монотонного характера этой зависимости.

Таким образом, для нахождения МП-оценки $\widehat{\theta}_{МП}$ следует

- найти решения уравнения правдоподобия (или системы уравнений) правдоподобия

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{\partial \ln L(\theta)}{\partial \theta} = 0; \tag{10}$$

при этом оценкой $\widehat{\theta}_{МП}$ будет лишь то решение уравнения (10), которое зависит от $x_1, x_2, \ldots, x_n$;

- среди решений, лежащих внутри множества значений неизвестного параметра θ, выделить точки максимума;
- если уравнение (система уравнений) (10) не определено, не разрешимо или среди решений нет точки максимума внутри области значений неизвест-

ного параметра θ, то точку максимума следует искать на границе этой области.

Отметим, что при достаточно общих условиях регулярности, накладываемых на изучаемый закон распределения $p(x;\theta)$, МП-оценки являются состоятельными, асимптотически несмещенными и эффективными.

<u>Пример 3.</u> Исследуемая случайная величина X распределена по закону Пуассона с неизвестным значением параметра λ. Найти МП-оценку этого параметра по независимой выборке $\{x_1, x_2, \ldots, x_n\}$ объема n.

<u>Решение.</u> Для случайной величины $X \sim \Pi(\lambda)$ имеем

$$p(x;\lambda) = P\{X = x;\lambda\} = \frac{\lambda^x}{x!} e^{-\lambda},\ x = 0,1,2,\ldots,$$

λ – неизвестный параметр.

Функция правдоподобия равна:

$$L(\lambda) = \prod_{i=1}^{n} p(x_i;\lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \frac{\lambda^{x_2}}{x_2!} e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda}.$$

Логарифмическая функция правдоподобия:

$$l(\lambda) = \ln L(\lambda) = (\sum_{i=1}^{n} x_i) \ln \lambda - \ln(x_1! \cdots x_n!) - n\lambda.$$

Уравнение правдоподобия:

$$\frac{\partial l(\lambda)}{\partial \lambda} = \sum_{i=1}^{n} x_i \frac{1}{\lambda} - n = 0,$$

отсюда $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$.

Легко видеть, что эта оценка несмещенная, так как

$$M\hat{\lambda}_{МП} = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} Mx_i = \frac{n\lambda}{n} = \lambda$$

(здесь все $x_i \sim \Pi(\lambda)$, $Mx_i = \lambda$, $i = 1,\ldots,n$).

Вычислим дисперсию оценки $\hat{\lambda}_{МП}$:

$$D\hat{\lambda}_{МП} = D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Dx_i = \frac{n\lambda}{n^2} = \frac{\lambda}{n}.$$

Так как $D\hat{\lambda}_{МП} \to 0$ при $n \to \infty$, то можно считать оценку $\hat{\lambda}_{МП}$ и эффективной.

<u>Пример 4.</u> Найти МП-оценки параметров a и σ^2 нормального распределения по независимой выборке $\{x_1, x_2, \ldots, x_n\}$ объема n.

<u>Решение.</u> Пусть независимая выборка $\{x_1, x_2, \ldots, x_n\}$ объема n извлечена из нормальной генеральной совокупности, т.е. исследуемая случайная вели-

чина X распределена по нормальному закону с математическим ожиданием $MX = a$, дисперсией $DX = \sigma^2 = D$ (значения этих параметров неизвестны до получения выборки), и имеет плотность

$$p(x;a,D) = \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x-a)^2}{2D}}.$$

Найдем функцию правдоподобия

$$L(a,D) = \prod_{i=1}^{n} p(x_i;a,D) = \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_1-a)^2}{2D}} \cdot \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_2-a)^2}{2D}} \cdots \frac{1}{\sqrt{2\pi D}} e^{-\frac{(x_n-a)^2}{2D}} =$$

$$= \left(\frac{1}{2\pi D}\right)^{n/2} e^{-\frac{1}{2D}\sum_{i=1}^{n}(x_i-a)^2}.$$

Соответствующая логарифмическая функция правдоподобия имеет вид

$$l(a,D) = \ln L(a,D) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln D - \frac{1}{2D}\sum_{i=1}^{n}(x_i - a)^2.$$

Дифференцируя $l(a,D)$ по a и D и последовательно приравнивая соответствующие частные производные к нулю, получаем систему уравнений правдоподобия:

$$\begin{cases} \dfrac{\partial l(a,D)}{\partial a} = \dfrac{1}{D}\sum_{i=1}^{n}(x_i - a) = 0, \\ \dfrac{\partial l(a,D)}{\partial D} = -\dfrac{n}{2D} + \dfrac{1}{2D^2}\sum_{i=1}^{n}(x_i - a)^2 = 0. \end{cases}$$

Решение этой системы относительно a и D дает оценки максимального правдоподобия этих параметров

$$\widehat{a}_{МП} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x} \text{ и } \widehat{D}_{МП} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = D_в.$$

Таким образом, МП-оценками неизвестного математического ожидания и неизвестной дисперсии являются выборочное среднее и выборочная дисперсия соответственно.

Проверим, будут ли найденные оценки несмещенными.

В общем случае все x_i, составляющие выборку, распределены по тому же закону, что и случайная величина X, т.е. $x_i \sim N(a,\sigma^2)$, поэтому $Mx_i = a,\ Dx_i = \sigma^2 = D$ для всех $i = 1,\ldots,n.$

Найдем $M\widehat{a}_{МП}$, используя свойства математического ожидания:

$$M\widehat{a}_{МП} = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} Mx_i = \frac{na}{n} = a.$$

Так как математическое ожидание оценки равно оцениваемому параметру, то МП-оценка математического ожидания в виде выборочного среднего является несмещенной. Используя свойства дисперсии, найдем дисперсию $\hat{a}_{МП}$:

$$D\hat{a}_{МП} = D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Dx_i = \frac{nD}{n^2} = \frac{D}{n} \to 0 \text{ при } n \to \infty.$$

Прежде чем определить $M\hat{D}_{МП}$, представим МП-оценку неизвестной дисперсии в виде

$$\hat{D}_{МП} = \frac{1}{n}\sum_{i=1}^{n}(x_i - a + a - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 + (\bar{x} - a)^2 - 2(\bar{x} - a)\cdot$$

$$\cdot\frac{1}{n}\sum_{i=1}^{n}(x_i - a) = \frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\bar{x} - a)^2.$$

Найдем $M\hat{D}_{МП}$:

$$M\hat{D}_{МП} = M\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - a)^2 - (\bar{x} - a)^2\right] = \frac{1}{n}\sum_{i=1}^{n} M(x_i - a)^2 - M(\bar{x} - a)^2 =$$

$$= \frac{1}{n}nD - \frac{D}{n} = D - \frac{D}{n} \neq D.$$

Так как $M\hat{D}_{МП} \neq D$, то МП-оценка неизвестной дисперсии, найденная в виде выборочной дисперсии, является смещенной, хотя, конечно же, асимптотическая несмещенность имеет место; смещение оценки равно $-\frac{D}{n}$, при увеличении объема выборки, т.е. при $n \to \infty$, смещение стремится к нулю.

Обычно смещение в оценке $\hat{D}$ устраняют, следуя специальной методике. Несмещенной и асимптотически эффективной оценкой дисперсии будет так называемая исправленная выборочная дисперсия

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

Она действительно будет несмещенной оценкой теоретической дисперсии, так как

$$Ms^2 = M\left(\frac{n}{n-1}D_в\right) = \frac{n}{n-1}MD_в = \frac{n}{n-1}M\hat{D}_{МП} = \frac{n}{n-1}\left(D - \frac{D}{n}\right) = D.$$

Таким образом, несмещенными оценками неизвестного математического ожидания и неизвестной дисперсии нормальной случайной величины будут

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$\hat{D} = \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2.$$

Задачи

21. Для случайной величины *X*, распределенной по экспоненциальному (показательному) закону с плотностью $p(x;\lambda) = \lambda e^{-\lambda x}$, $x \geq 0$, найти МП-оценку параметра λ по независимой выборке $\{x_1, x_2, \ldots, x_n\}$ и исследовать ее свойства.

22. По независимой выборке $\{x_1, x_2, \ldots, x_n\}$ найти МП-оценку неизвестного параметра α случайной величины *X*, распределенной по закону Максвелла с плотностью вероятности $p(x;\alpha) = \sqrt{\frac{2}{\pi}} \frac{x^2}{\alpha^3} e^{-\frac{x^2}{2\alpha^2}}$.

23. Продолжительность безотказной работы датчика является случайной величиной с плотностью распределения $p(t;\alpha) = \frac{1}{\alpha} e^{-\frac{t}{\alpha}}$, $(t \geq 0, \alpha \geq 0)$. По фиксированным значениям $t_1, t_2, \ldots, t_n$ найти МП-оценку неизвестного параметра α.

24. По независимой выборке объема n, извлеченной из генеральной совокупности, имеющей геометрическое распределение $P\{X = k\} = q^{k-1} p$, оценить неизвестный параметр p.

25. Случайная величина *X* представляет собой количество срывов поставок потребителям фирмами, производящими однородную продукцию. За определенный период обследовано 7 фирм, у которых количество срывов поставок соответственно равно: 6, 3, 1, 3, 4, 0, 2. Полагая, что случайная величина *X* распределена по закону Пуассона, найти МП-оценку параметра λ.

26. Случайная величина *X* подчинена биномиальному закону распределения с неизвестным параметром p. Дано эмпирическое распределение X:

x_i	0	1	2	3	4	5	6	7
n_i	2	3	10	22	26	20	12	5

Найти МП-оценку параметра p.

27. Проведено 10 независимых наблюдений случайной величины *X* – дохода на душу населения (в денежных единицах):

x_i	3	5	7	8	10
n_i	8	10	45	25	12

Считая, что X – распределена по закону Парето с плотностью

$$p(x,a)=\begin{cases}\dfrac{a}{3}\left(\dfrac{3}{x}\right)^{a+1}, & x\ge 3,\ a>0;\\ 0, & x<3\end{cases},$$

Найти МП – оценку неизвестного параметра a.

2.2. Метод моментов

Пусть независимая выборка $\{x_1,x_2,\ldots,x_n\}$ извлечена из распределения с плотностью $p(x;\theta_1,\theta_2,\ldots,\theta_r)$, зависящей от r неизвестных параметров $\theta_1,\theta_2,\ldots,\theta_r$. Предположим, что первые r теоретические начальные моменты существуют и конечны:

$$m_k(\theta_1,\theta_2,\ldots,\theta_r)=MX^k=\int x^k p(x;\theta_1,\theta_2,\ldots,\theta_r)dx,\quad k=1,\ldots,r.$$

(для дискретного распределения интеграл следует заменить суммой).

По выборке $\{x_1,x_2,\ldots,x_n\}$ построим так называемые выборочные или эмпирические начальные моменты $\widehat{m}_k$, которые будут несмещенными оценками соответствующих теоретических моментов:

$$\widehat{m}_k=\frac{1}{n}\sum_{i=1}^{n}x_i^k,\quad k=1,\ldots,r.$$

Метод моментов состоит в том, что оценки неизвестных параметров $\widehat{\theta}_k,\ k=1,\ldots,r,$ находятся как решение системы уравнений:

$$\begin{cases}m_1(\theta_1,\ldots,\theta_r)=\widehat{m}_1,\\ m_2(\theta_1,\ldots,\theta_r)=\widehat{m}_2,\\ \ldots\ldots\ldots\ldots\ldots\ldots\\ m_r(\theta_1,\ldots,\theta_r)=\widehat{m}_r.\end{cases}\qquad(11)$$

Использование начальных моментов необязательно; здесь могут использоваться центральные теоретические и соответствующие им эмпирические моменты.

К достоинствам метода следует отнести его сравнительно простую вычислительную реализацию, а также то, что оценки, полученные из решения системы (11), являются функциями от выборочных моментов. Это упрощает исследование статистических свойств оценок метода моментов. В то же время оценки не всегда будут асимптотически эффективными, и в этом отношении они уступают оценкам, полученным методом максимального правдоподобия. Тем не менее, метод моментов часто очень удобен на практике.

Пример 5. Случайная величина $X \sim N(a,\sigma^2)$, при этом значения параметров a и σ^2 неизвестны. Найти методом моментов оценки этих параметров по независимой выборке $\{x_1, x_2, \ldots, x_n\}$ объема n.

Решение. Так как для $X \sim N(a,\sigma^2)$ первый и второй начальные теоретические моменты существуют и равны соответственно

$$m_1 = MX^1 = a, \quad m_2 = MX^2 = \sigma^2 + a^2,$$

то система (11) для определения оценок $\hat{a}$ и $\hat{\sigma}^2$ примет вид

$$\begin{cases} a = \dfrac{1}{n}\sum_{i=1}^{n} x_i, \\ \sigma^2 + a^2 = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2. \end{cases}$$

Решениями этой системы будут

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = D_в.$$

Мы получили методом моментов те же оценки неизвестного математического ожидания и неизвестной дисперсии, что и методом максимального правдоподобия.

Замечание. При применении метода моментов к группированным выборкам, т.е. выборкам, представленным в виде примыкающих друг к другу интервалов шириной h, необходима корректировка оценок теоретических моментов. Эмпирические моменты, найденные в этом случае по серединам интервалов, не всегда будут несмещенными оценками теоретических моментов. Смещение в оценках устраняют, вводя так называемые поправки Шеппарда.

Несмещенной оценкой первого теоретического начального момента будет

$$\hat{m}_1 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^*,$$

где x_i^* – середина i – го интервала, а n_i – соответствующая частота. Несмещенная оценка второго теоретического начального момента равна

$$\hat{m}_2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i^*)^2 - \frac{h^2}{12}.$$

Здесь величина $-\dfrac{h^2}{12}$ и есть поправка Шеппарда.

Несмещенные оценки третьего и четвертого теоретических начальных моментов с учетом поправок Шеппарда запишутся как

$$\hat{m}_3 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i^*)^3 - \frac{h^2}{4}\frac{1}{n}\sum_{i=1}^{k} n_i x_i^*,$$

$$\widehat{m}_4 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i^*)^4 - \frac{h^2}{2}\frac{1}{n}\sum_{i=1}^{k} n_i (x_i^*)^2 + \frac{7h^4}{240}.$$

Пример 6. При тестировании группы студентов есть основание считать, что средний балл X – это равномерно распределенная на отрезке $[a,b]$ случайная величина. Результаты обследования представлены в виде интервального вариационного ряда:

$x_{i-1} - x_i$	0 – 2	2 – 4	4 – 6	6 – 8	8 – 10
n_i	12	10	9	9	10

$n = \sum n_i = 50.$

Найти методом моментов оценки параметров $\widehat{a}$ и $\widehat{b}$.

Решение. Для равномерного на отрезке $[a,b]$ распределения имеем:

$$p(x;a,b) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}.$$

Теоретические начальные моменты первого и второго порядков равны соответственно:

$$m_1 = MX = \int_a^b xp(x;a,b)dx = \frac{1}{b-a}\int_a^b xdx = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2} = m_1(a,b);$$

$$m_2 = MX^2 = \int_a^b x^2 p(x;a,b)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3} = m_2(a,b).$$

Для нахождения эмпирических начальных моментов от заданного интервального ряда перейдем к точечному:

x_i^*	1	3	5	7	9
n_i	12	10	9	9	10

Тогда

$$\widehat{m}_1 = \frac{1}{n}\sum_{i=1}^{5} n_i x_i^* = \frac{1}{50}(12\cdot 1 + 10\cdot 3 + 9\cdot 5 + 9\cdot 7 + 10\cdot 9) = 4{,}8;$$

$$\widehat{m}_2 = \frac{1}{n}\sum_{i=1}^{5} n_i (x_i^*)^2 - \frac{h^2}{12} = \frac{1}{50}(12\cdot 1^2 + 10\cdot 3^2 + 9\cdot 5^2 + 9\cdot 7^2 + 10\cdot 9^2) - \frac{4}{12} =$$

$= 31{,}2267.$

По методу моментов оценки двух неизвестных параметров $\widehat{a}$ и $\widehat{b}$ находятся из решения системы уравнений:

$$\begin{cases} m_1(a,b) = \widehat{m}_1, \\ m_2(a,b) = \widehat{m}_2. \end{cases}$$

Имеем

$$\begin{cases} \dfrac{a+b}{2} = 4{,}8, \\ \dfrac{a^2 + ab + b^2}{3} = 31{,}2267. \end{cases}$$

Отсюда

$$\begin{cases} a+b = 9{,}6, \\ a^2 + ab + b^2 = 93{,}6801. \end{cases}$$

Из решения этой системы уравнений получаем $\widehat{a} \approx -0{,}31, \quad \widehat{b} \approx 9{,}76.$

Задачи

28. Случайная величина X распределена по закону Пуассона $P_{x_i}(\lambda) = \dfrac{\lambda^{x_i}}{x_i!} e^{-\lambda}$, где x_i – число появлений события в i - м испытании. Найти методом моментов по выборке $\{x_1, x_2, \ldots, x_n\}$ оценку параметра λ.

29. Найти методом моментов точечную оценку параметра p геометрического распределения $P\{X = x_i\} = (1-p)^{x_i - 1} p$, где x_i – число испытаний, проведенных до появления события, p - вероятность появления события в одном испытании.

30. Случайная величина X распределена по биномиальному закону с неизвестным параметром p. Приведено эмпирическое распределение числа появлений события в 10 опытах по 5 испытаний в каждом:

x_i	0	1	2	3	4
n_i	5	2	1	1	1

Найти ММ-оценку параметра p.

31. Случайная величина X имеет показательное распределение $p(x;\lambda) = \lambda e^{-\lambda x}$ $(x \geq 0)$. Приведено эмпирическое распределение среднего времени $n = 200$ элементов:

x_i	2,5	7,5	12,5	17,5	22,5	27,5
n_i	133	45	15	4	2	1

x_i - среднее время работы элемента в часах, n_i - количество элементов. Найти ММ-оценку параметра λ.

32. Случайная величина X подчинена равномерному закону распределения с параметрами a и b. Приведено эмпирическое распределение средней ошибки $n = 200$ измерений дальности:

x_i	3	5	7	9	11	13	15	17	19	21
n_i	21	16	15	26	22	14	21	22	18	25

Найти ММ-оценки параметров a и b.

33. Случайная величина X распределена по «двойному» закону Пуассона:

$$P\{X = x_i\} = \frac{1}{2}\frac{\lambda_1^{x_i} e^{-\lambda_1}}{x_i!} + \frac{1}{2}\frac{\lambda_2^{x_i} e^{-\lambda_2}}{x_i!}.$$

Приведено эмпирическое распределение числа появлений события в $n = 327$ испытаниях

x_i	0	1	2	3	4	5	6	7	8	9	10
n_i	28	47	81	67	53	94	13	8	3	2	1

Найти методом моментов оценки параметров λ_1 и λ_2.

34. Полагая, что рост мужчин есть случайная величина $X \sim N(a,\sigma)$. Найти методом моментов оценки параметров a и σ по следующим данным

Рост, см.	Число мужчин
143 – 146	1
146 – 149	2
149 – 152	8
152 – 155	26
155 – 158	65
158 – 161	120
161 – 164	181
164 – 167	201
167 – 170	170
170 – 173	120
173 – 176	64
176 – 179	28
179 – 182	10
182 – 185	3
185 – 188	1

2.3. Интервальное оценивание

В практических задачах наряду с точечными оценками необходимо указывать интервалы, которые бы с практической достоверностью, т.е. с вероятностью, близкой к единице, покрывали истинное неизвестное значение параметра. Такие интервалы называются доверительными или интервальными оценками, а вероятность, с которой доверительный интервал покрывает неизвестное значение параметра, называется доверительной вероятностью или надежностью.

В математической статистике доверительные интервалы используются для определения точности оценки $\hat{\theta}$ неизвестного параметра θ, а доверительные вероятности – для определения надежности. Интервальное оценивание особенно необходимо при малом числе наблюдений, когда точечная оценка малонадежна.

Приведем правила определения границ доверительных интервалов для параметров нормального распределения и неизвестной вероятности успеха (т.е. появления интересующего нас события) в условиях n испытаний Бернулли.

Пусть случайная величина X распределена по нормальному закону, параметрами которого являются $a = MX$ и $D = \sigma^2 = DX$.

Доверительный интервал для неизвестного математического ожидания при известной дисперсии σ^2 имеет вид

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}}, \tag{12}$$

где $\bar{x}$ – выборочное среднее, являющееся несмещенной оценкой неизвестного математического ожидания; n – объем выборки; σ – среднее квадратическое отклонение, значение которого известно, t_γ – находится из уравнения $\Phi(t_\gamma) = \frac{\gamma}{2}$, где γ – заданная доверительная вероятность, а $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ – функция Лапласа (см. приложение 3). Значения доверительной вероятности стандартизированы: $\gamma = 0{,}9;\ 0{,}95;\ 0{,}99;\ 0{,}999$.

Величина $\Delta = t_\gamma \frac{\sigma}{\sqrt{n}}$ характеризует точность оценивания. Объем выборки, обеспечивающий заданную точность Δ, определяется как

$$n = \left(\frac{t_\gamma \sigma}{\Delta} \right)^2. \tag{13}$$

Доверительный интервал для неизвестного математического ожидания при неизвестной дисперсии σ^2 записывается в виде

$$\bar{x} - t_\gamma \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{s}{\sqrt{n}}. \qquad (14)$$

Сравнивая (14) с (12), легко убедиться в том, что доверительные границы здесь определяются аналогичным образом, за исключением того, что вместо σ используется оценка $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, а коэффициент t_γ находится из таблиц квантилей распределения Стьюдента по заданной доверительной вероятности γ и числу степеней свободы $k = n-1$ (см. приложение 5), $t_\gamma = t\left(\frac{1+\gamma}{2}, n-1\right)$.

<u>Доверительный интервал для неизвестной дисперсии</u> $\sigma^2 = D$ имеет вид

$$\frac{s^2(n-1)}{U_2} < D < \frac{s^2(n-1)}{U_1}, \qquad (15)$$

где $s^2 = \widehat{D}$ – несмещенная оценка дисперсии, найденная по выборке объема n; $U_2 = \chi^2\left(\frac{1+\gamma}{2}; n-1\right)$ и $U_1 = \chi^2\left(\frac{1-\gamma}{2}; n-1\right)$ находятся из таблиц квантилей распределения χ^2 («хи-квадрат») по соответствующей вероятности γ и числу степеней свободы $k = n-1$ (см. приложение 4).

<u>Доверительный интервал для неизвестной вероятности успеха p в условиях n испытаний Бернулли</u> записывается как

$$\widehat{p} - t_\gamma \sqrt{\frac{\widehat{p}\widehat{q}}{n}} < p < \widehat{p} + t_\gamma \sqrt{\frac{\widehat{p}\widehat{q}}{n}}, \qquad (16)$$

где $\widehat{p} = \frac{m}{n}$ – относительная частота появления события A, $\widehat{q} = 1 - \widehat{p}$, t_γ – коэффициент Стьюдента. Из таблиц квантилей t – распределения (см. приложение 5) этот коэффициент находится как квантиль уровня $\frac{1+\gamma}{2}$, т.е. $t_\gamma = t\left(\frac{1+\gamma}{2}, n-1\right)$.

<u>Пример 7.</u> Из большой партии изделий отобрано наугад для контроля 500 штук, причем среди них 20 не удовлетворяют стандарту. Найти с доверительной вероятностью 0,95 интервал, содержащий процент брака во всей партии.

<u>Решение.</u> Для оценки вероятности брака используем неравенство (16). Здесь $\widehat{p} = \frac{m}{n} = \frac{20}{500} = 0{,}04$, $\widehat{q} = 1 - \widehat{p} = 1 - 0{,}04 = 0{,}96$, $t_\gamma = t_{0,95} = \; = 1{,}96$ (см.

приложение 5), тогда $0{,}023 < p < 0{,}057$ или процент брака во всей партии будет от 2,3% до 5,7%.

Пример 8. Фирма коммунального хозяйства на основе выборки оценивает среднюю квартплату за квартиры определенного типа с надежностью не менее 99% и погрешностью, меньшей 10 д.е. Предполагая, что квартплата имеет нормальное распределение со средним квадратическим отклонением, не превышающим 35 д.е., найти минимальный объем выборки.

Решение. По условию требуется найти такое n, при котором $P\{|\bar{x}-a|<10\} \geq 0{,}99$. Приравняв $1-\alpha = 0{,}99$, из таблицы значений функции Лапласа (приложение 3) найдем $t_{0,99}$: $\Phi(t_{0,99}) = \dfrac{0{,}99}{2} = 0{,}495$; $t_{0,99} = 2{,}6$. При $\Delta = 10$ и $\sigma = 35$ из формулы (13) получим

$$n = \frac{t_{0,99}^2 \sigma^2}{\Delta^2} = \frac{6{,}76 \cdot 1225}{100} = 82{,}81.$$

Так как с ростом $1-\alpha$ и уменьшением Δ растет n, то $n \geq 82{,}81$ и тогда минимальный объем выборки будет равен $n_{\min} = 83$.

Пример 9. Признак X генеральной совокупности распределен нормально. Произведена выборка, данные которой приведены в таблице

x_i	0,1	0,2	0,3	0,4	0,5
n_i	2	4	7	6	1

Найти выборочное и исправленное среднее квадратическое отклонение (σ_B и s), а также доверительный интервал, покрывающий неизвестное среднее квадратическое отклонение σ с вероятностью $\gamma = 0{,}95$.

Решение. Сначала найдем

$$\bar{x} = \frac{1}{20}\sum_{i=1}^{5} x_i n_i = \frac{0{,}1\cdot 2 + 0{,}2\cdot 4 + 0{,}3\cdot 7 + 0{,}4\cdot 6 + 0{,}5\cdot 1}{20} = 0{,}3,$$

$$\sigma_B{}^2 = \frac{1}{20}\sum_{i=1}^{5}(x_i - 0{,}3)^2 n_i = \frac{(0{,}1-0{,}3)^2\cdot 2 + (0{,}2-0{,}3)^2\cdot 4 + (0{,}3-0{,}3)^2\cdot 7 +}{20}$$

$$\frac{+(0{,}4-0{,}3)^2\cdot 6 + (0{,}5-0{,}3)^2\cdot 1}{} = 0{,}011, \text{ отсюда } \sigma_B = 0{,}1049,$$

$$s^2 = \frac{n}{n-1}\sigma_B{}^2 = \frac{20}{19}0{,}011 = 0{,}0116, \quad \text{тогда } s = 0{,}1076.$$

Из таблиц квантилей распределения χ^2 (см. приложение 4) определим значения

$$U_1 = \chi^2\left(\frac{1-\gamma}{2}; n-1\right) = \chi^2(0{,}025;19) = 8{,}91,$$

$$U_2 = \chi^2\left(\frac{1+\gamma}{2}; n-1\right) = \chi^2(0{,}975;19) = 32{,}9.$$

С вероятностью 0,95 согласно (15) будем иметь

$$\frac{19 \cdot 0{,}0116}{32{,}9} < \sigma^2 < \frac{19 \cdot 0{,}0116}{8{,}91}.$$

Отсюда с доверительной вероятностью 0,95 можно утверждать, что среднее квадратическое отклонение σ будет находиться в интервале $0{,}081 < \sigma < 0{,}157$.

Пример 10. С целью определения средней суммы вкладов в банке, имеющем 2200 вкладчиков, произведено выборочное обследование 111 вкладов, результаты которого даны в таблице

Средняя сумма вклада, тыс.руб., x_i	10-30	30-50	50-70	70-90	90-110	110-130
Число вкладов, n_i	1	3	10	30	60	7

Найти доверительный интервал для неизвестного математического ожидания, который можно было бы гарантировать с вероятностью 0,95.

Решение. Сначала найдем выборочное среднее и выборочную дисперсию, для чего в качестве значений признака X берем середины интервалов. Тогда

$$\bar{x} = \frac{1}{111}\left(20 \cdot 1 + 40 \cdot 3 + 60 \cdot 10 + 80 \cdot 30 + 100 \cdot 60 + 120 \cdot 7\right) = 89{,}91,$$

$$\sigma_B{}^2 = \frac{1}{111}\Big[(20-89{,}91)^2 \cdot 1 + (40-89{,}91)^2 \cdot 3 + (60-89{,}91)^2 \cdot 10 +$$

$$+ (80-89{,}91)^2 \cdot 30 + (100-89{,}91)^2 \cdot 60 + (120-89{,}91)^2 \cdot 7\Big] = 330{,}64.$$

Так как $\Phi(t_\gamma) = \dfrac{\gamma}{2}$, то $\Phi(t_\gamma) = \dfrac{0{,}95}{2} = 0{,}475 \quad \Rightarrow \quad t_\gamma = 1{,}96$ и по формуле (12) имеем

$$89{,}91 - 1{,}96\frac{\sqrt{330{,}64}}{\sqrt{111}} < a < 89{,}91 + 1{,}96\frac{\sqrt{330{,}64}}{\sqrt{111}}.$$

Окончательно $86{,}53 < a < 93{,}29$.

Здесь мы использовали формулу интервального оценивания неизвестного математического ожидания, где коэффициент t_γ определялся из таблиц функции Лапласа. Так как фактически дисперсия случайной величины X оценивалась по заданной выборке, то, строго говоря, при определении t_γ следовало бы воспользоваться распределением Стьюдента. Но для выборки большого объема (n = 111) различия между этими коэффициентами нет.

Задачи

35. Из многочисленного коллектива работников фирмы случайным образом отобрано $n = 25$ человек. Средняя заработная плата этих работников составила $\bar{x} = 700$ д.е. при среднем квадратическом отклонении $s = 100$ д.е. Требуется найти с доверительной вероятностью $\gamma = 0{,}95$ интервальную оценку для:

а) средней месячной заработной платы на фирме,

б) суммы затрат на заработную плату отдела, состоящего из 520 сотрудников.

36. При анализе точности фасовочного автомата было произведено $n = 24$ контрольных взвешиваний пятисотграммовых пачек кофе. По результатам измерений рассчитано среднее квадратическое отклонение $\sigma = 0{,}8$ г. Требуется с доверительной вероятностью $\gamma = 0{,}95$, оценить точность фасовочного автомата, т.е. определить интервальную оценку для σ.

37. С помощью случайного отбора определяется стаж работы служащих фирмы. Предполагается, что он подчиняется нормальному закону. Каким должен быть объем выборки, чтобы с доверительной вероятностью 0,95 можно было утверждать, что, принимая полученный средний стаж работы за истинный, совершается погрешность, не превышающая 0,5 года, если $\sigma = 2{,}7$ года?

38. В таблице приведены результаты выборочного обследования 100 рабочих фирм с целью определения времени, затрачиваемого на обработку детали

Время обработки, мин. x_i	3,6-4,2	4,2-4,8	4,8-5,4	5,4-6	6-6,6
Число рабочих, n_i	14	33	35	12	6

Требуется найти: выборочное среднее, дисперсию, среднее квадратическое отклонение и границы, в которых с надежностью 0,95 заключено среднее время обработки детали рабочим.

39. Произведено выборочное измерение выработки на земляных работах 145 рабочих. В результате этого обследования средняя выработка определена $4{,}95\,м^3$ на одного рабочего, а среднее квадратическое отклонение равно $1{,}5\,м^3$. Найти доверительные границы для генерального среднего (неизвестного математического ожидания) с вероятностью $\gamma = 0{,}99$.

40. Выборочно обследовали качество кирпича. Из 1600 проб в 32 случаях кирпич оказался бракованным. Требуется определить, в каких пределах заключается доля брака всей продукции, если результат необходимо гарантировать с вероятностью $\gamma = 0{,}95$.

41. Определить численность выборки при обследовании остатков на расчетных счетах у клиентов национального банка, чтобы с вероятностью 0,683 ошибка репрезентативности не превышала 5 руб. и $\sigma = 120$ руб.

42. В таблице приведены результаты выборочного обследования заработной платы 100 рабочих предприятия

Заработная плата, у.е.	100-120	120-140	140-160	160-180	180-200
Число рабочих	17	40	32	8	3

Требуется найти границы, в которых с надежностью 0,95 заключена заработная плата рабочих предприятия.

43. Продукция двух фирм была проанализирована и оказалась однородной по составу. На предприятиях этих фирм производятся одинаковые товары. На одной из фирм (A) провели частичную модернизацию, на другой (B) нет. Через месяц после модернизации производства фирмы A была произведена случайная выборка по 50 изделий обеих фирм. При этом оценивалось качество изделий (в баллах). Средние баллы и выборочные дисперсии оказались следующими: $\bar{x}_A = 323$, $\bar{x}_B = 297$, $s_A^2 = 441$, $s_B^2 = 529$. За меру эффекта воздействия модернизации на качество продукции взять разность средних. Найти для разности средних 99%-й доверительный интервал.

Контрольные и тестовые задания

1. Оценка $\hat{\theta}$ неизвестного параметра θ является:

а) случайной величиной, близкой в некотором смысле к истинному значению оцениваемого параметра;

б) неслучайной величиной, близкой в определенном смысле к истинному значению оцениваемого параметра;

в) случайной или неслучайной величиной, приближенно определяющей истинное значение оцениваемого параметра;

г) величиной, не зависящей от выборки и приближенно определяющей значение оцениваемого параметра.

2. Если для оценки $\hat{\theta}$ неизвестного параметра θ выполняется $\lim\limits_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0$ при любых $\varepsilon > 0$, то эта оценка называется:

а) несмещенной;
б) эффективной;
в) состоятельной;
г) несостоятельной;
д) смещенной.

3. Оценка $\widehat{\theta}$ неизвестного параметра θ называется несмещенной, если:

а) $\lim\limits_{n\to\infty} P\left(\left|\widehat{\theta}-\theta\right|<\varepsilon\right)=1$;

б) $M\left(\widehat{\theta}\right)=\theta$;

в) среди всех прочих оценок того же параметра она обладает наименьшей дисперсией;

г) $\lim\limits_{n\to\infty} P\left(\left|\widehat{\theta}-\theta\right|>\varepsilon\right)=0$;

д) $M\left(\widehat{\theta}\right)\neq\theta$.

4. Для выборки $\{x_1, x_2, \dots, x_n\}$, извлеченной из генеральной совокупности, распределение которой описывается функцией плотности $p(x,\theta)$, соотношение $L(\theta)=\prod\limits_{i=1}^{n} p(x_i,\theta)$ называется:

а) логарифмической функцией правдоподобия;
б) уравнением правдоподобия;
в) эмпирической функцией распределения;
г) функцией правдоподобия.

5. Какой оценкой является выборочная дисперсия $D_B=\dfrac{1}{n}\sum\limits_{i=1}^{n}\left(x_i-\overline{x}\right)^2$?

а) несмещенной;
б) смещенной;
в) состоятельной;
г) несостоятельной.

6. Выборочная средняя $\overline{x}=\dfrac{1}{n}\sum\limits_{i=1}^{n}x_i$, как точечная оценка неизвестного математического ожидания, не является:

а) состоятельной;
б) эффективной;
в) несмещенной;
г) смещенной.

7. Доверительным интервалом для неизвестного параметра θ называется интервал $\left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)$, который:

а) с заданной вероятностью покрывает истинное значение параметра θ;
б) содержит истинное значение параметра θ с вероятностью 1;
в) с заданной вероятностью не содержит истинное значение параметра θ;

г) содержит точечную оценку этого параметра.

8. Если $\left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)$ доверительный интервал для параметра θ надежности γ, то:

а) $P(\theta \notin \left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)) = \gamma$; б) $P(\theta \in \left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)) = 1-\gamma$;

в) $P(\theta \in \left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)) < \gamma$; г) $P(\theta \in \left(\widehat{\theta}-\Delta, \widehat{\theta}+\Delta\right)) = \gamma$;

д) $P(\left|\widehat{\theta}-\theta\right| \geq \Delta) > 1-\gamma$.

9. В интервальной оценке $\overline{x}-\Delta < \alpha < \overline{x}+\Delta$ математического ожидания нормального распределения с известной дисперсии σ^2 величина Δ (точность оценивания) равна:

а) $\Delta = t_\gamma \cdot \dfrac{\sigma}{n}$; б) $\Delta = t_\gamma \cdot \dfrac{\sigma^2}{\sqrt{n}}$; в) $\Delta = t_\gamma \cdot \dfrac{\sigma}{\sqrt{n}}$;

г) $\Delta = t_\gamma \cdot \dfrac{\sqrt{n}}{\sigma}$; д) $\Delta = t_\gamma \cdot \dfrac{s}{\sqrt{n}}$,

где t_γ определяется из условия $\Phi\left(t_\gamma\right) = \dfrac{\gamma}{2}$.

10. Доверительный интервал надежности γ для математического ожидания a нормальной генеральной совокупности с неизвестной дисперсией определяется как:

а) $\left(\overline{x}-t_\gamma \cdot \dfrac{\sigma}{\sqrt{n}},\quad \overline{x}+t_\gamma \cdot \dfrac{\sigma}{\sqrt{n}}\right)$; б) $\left(\overline{x}-t_\gamma \cdot \dfrac{s}{\sqrt{n}},\quad \overline{x}+t_\gamma \cdot \dfrac{s}{\sqrt{n}}\right)$;

в) $\left(\overline{x}-t_\gamma \cdot \dfrac{s}{n},\quad \overline{x}+t_\gamma \cdot \dfrac{s}{n}\right)$; г) $\left(\overline{x}-t_\gamma \cdot \dfrac{\sigma}{n},\quad \overline{x}+t_\gamma \cdot \dfrac{\sigma}{n}\right)$,

где $t_\gamma = t\left(\dfrac{1+\gamma}{2}, n-1\right)$ – квантиль уровня $\dfrac{1+\gamma}{2}$ распределения Стьюдента с числом степеней свободы $n-1$.

11. Как изменяется точность оценки математического ожидания нормального распределения с ростом объема выборки?

а) улучшается;

б) ухудшается;

в) не изменяется;

г) может, как улучшаться, так и ухудшаться.

12. Какое из неравенств определяет доверительный интервал надежности γ для дисперсии нормально распределительной генеральной совокупности при неизвестном математическом ожидании?

а) $\dfrac{(n-1)s^2}{u_1} < \sigma^2 < \dfrac{(n-1)s^2}{u_2}$; б) $\dfrac{(n-1)s^2}{u_2} < \sigma^2 < \dfrac{(n-1)s^2}{u_1}$;

в) $\dfrac{n \cdot s^2}{u_2} < \sigma^2 < \dfrac{n \cdot s^2}{u_1}$; г) $\dfrac{s\sqrt{n-1}}{\sqrt{u_2}} < \sigma < \dfrac{s \cdot \sqrt{n}}{\sqrt{u_1}}$,

где $u_1 = \chi^2\left(\dfrac{1-\gamma}{2}; n-1\right)$, $u_2 = \chi^2\left(\dfrac{1+\gamma}{2}; n-1\right)$ соответственно квантили уровней $\dfrac{1-\gamma}{2}$ и $\dfrac{1+\gamma}{2}$ распределения χ^2 с числом степеней свободы $n-1$.

Глава 3. Статистическая проверка гипотез

Основными составными частями математической статистики являются теория статистического оценивания неизвестных значений параметров, участвующих в описании анализируемой модели, и теория проверки на основе выборочных наблюдений различных предположений о параметрах или природе анализируемой модели. Такие предположения носят статистический характер (или касаются вероятностных свойств анализируемого объекта) и называются статистическими гипотезами. По своему прикладному содержанию высказываемые в ходе статистической обработки данных гипотезы подразделяются на следующие типы:

- о числовых значениях параметров исследуемой генеральной совокупности;
- об общем виде закона распределения исследуемой случайной величины;
- об однородности двух или нескольких обрабатываемых выборок;
- об общем виде зависимости, существующей между анализируемыми показателями;
- о независимости и стационарности ряда наблюдений.

В данной главе мы рассмотрим основные понятия и постановку задачи, дадим общую логическую схему проверки статистических гипотез, а также приведем критерии проверки гипотез о параметрах нормальной генеральной совокупности и о виде закона распределения анализируемого признака.

3.1. Основные понятия и постановка задачи проверки гипотез

<u>Статистической гипотезой</u> называется предположение относительно параметров или вида распределения случайной величины. Гипотеза называется <u>параметрической</u>, если она касается значений параметров распределения известного вида. В <u>непараметрической</u> гипотезе речь идет об утверждении относительно вида распределения.

Гипотеза называется <u>простой</u>, если она однозначно определяет либо значение неизвестного параметра, либо распределение случайной величины, в противном случае гипотеза <u>сложная.</u> Например, простой гипотезой является предположение о том, что наблюдаемая случайная величина $X \sim N(0,1)$. Если же высказывается предположение, что $X \sim N(a,1)$, где $a_1 \le a \le a_2$, то это сложная гипотеза.

Та гипотеза, относительно которой ведется проверка, называется <u>основной</u> или <u>нулевой</u> гипотезой и обозначается H_0. Наряду с гипотезой H_0 рассматривается конкурирующая или альтернативная гипотеза H_1, которая должна быть принята в случае отклонения H_0. Например, если простая гипотеза $H_0 : \theta = \theta_0$ (гипотеза о равенстве параметра θ некоторому предполагаемому или гипотетическому значению θ_0), то в качестве альтернативной гипотезы можно рассмотреть $H_1 : \theta > \theta_0$, либо $H_1 : \theta < \theta_0$, либо $H_1 : \theta \neq \theta_0$ или, наконец, $H_1 : \theta = \theta_1$, где $\theta_1 \neq \theta_0$. Выбор альтернативной гипотезы определяется конкретным содержанием задачи.

Критерием проверки гипотезы называется правило, по которому принимается решение принять или отклонить гипотезу H_0. Решение здесь выносится в зависимости от значения специальным образом сконструированной случайной величины – статистики критерия или просто критерия, распределение которой известно и затабулировано. Множество значений критерия разбивается на две непересекающиеся области: область принятия нулевой гипотезы d_0, при попадании в которую нулевая гипотеза принимается, и критическую область d_1, при попадании в которую нулевая гипотеза H_0 отвергается и принимается альтернативная гипотеза H_1. Точка, разделяющая эти две области, называется критической или пороговым значением критерия.

В процессе проверки гипотезы можно либо прийти к правильному решению: принять гипотезу H_0, когда она на самом деле верна, или отвергнуть H_0, когда она на самом деле не верна; либо совершить одну из двух ошибок: отвернуть гипотезу H_0, когда она верна (ошибка первого рода или ошибка типа пропуска цели), или принять гипотезу H_0, когда она не верна, а верна конкурирующая гипотеза H_1 (ошибка второго рода или ошибка типа ложной тревоги). Последствия этих ошибок часто оказываются совершенно различными, поэтому желательно провести проверку гипотез таким образом, чтобы свести к минимуму вероятности ошибок обоих типов.

Обозначим через α вероятность ошибки первого рода: $\alpha = P(d_1 / H_0)$, β – вероятность ошибки второго рода: $\beta = P(d_0 / H_1)$. Вероятность отвергнуть неправильную гипотезу $P(d_1 / H_1) = 1 - \beta$ называют мощностью критерия, а α - уровнем значимости.

Критерий называется наиболее мощным, если из всех возможных критериев с заданным уровнем значимости α он обладает наибольшей мощностью, т.е. если его критическая область d_1^* такова, что $P(d_1^* / H_1) = \max\limits_{d_1} P(d_1 / H_1)$, где максимум берется по тем областям d_1, для которых $P(d_1 / H_0) = \alpha$.

Так как мощность критерия равна $1 - \beta$, то использование наиболее мощного критерия гарантирует при заданной вероятности ошибки первого рода α наименьшую, по сравнению с другими критериями, вероятность ошибки второго рода β.

3.2. Общая логическая схема проверки статистических гипотез

По своему назначению и характеру решаемых задач статистические критерии чрезвычайно разнообразны. Однако их объединяет общность логической схемы, по которой они строятся. Коротко эту схему можно описать так.

1. Выдвигается основная гипотеза H_0.

Если гипотеза параметрическая, то наряду с H_0 выдвигается конкурирующая гипотеза H_1, которая должна быть принята в случае отклонения H_0. Если же гипотеза H_0 касается вида закона распределения вероятностей, то H_1 формально не определяется: она состоит в отклонении H_0.

2. Задается уровень значимости α.

Для удобства значения α стандартизированы и принимаются обычно равными 0,1; 0,05; 0,01; 0,005; 0,001.

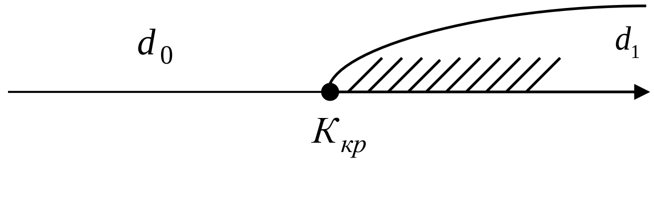
3. Выбирается критерий проверки гипотезы H_0.

Статистика критерия $K = K(x_1, x_2, \ldots, x_n)$, являясь функцией выборки, будет случайной величиной, закон распределения которой известен и затабулирован. Чаще всего в качестве таких известных распределений используют $N(0,1), t-, \chi^2-, F-$распределения.

4. Из таблиц распределения критерия по заданному уровню значимости α выбирается критическая точка $K_{кр}$, которая делит множество значений критерия на область принятия нулевой гипотезы d_0 и критическую область d_1.

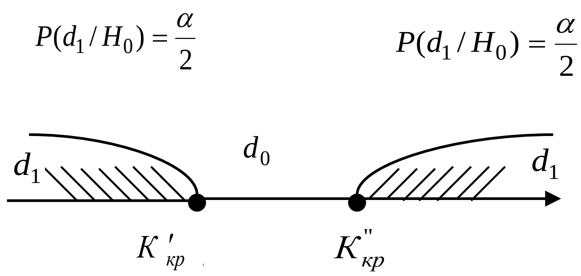
«Размер» критической области определяется уровнем значимости α, «положение» области на оси определяется видом конкурирующей гипотезы H_1.

Так, если $H_0: \theta = \theta_0$, $H_1: \theta = \theta_1$, где $\theta_1 > \theta_0$, либо $H_1: \theta > \theta_0$, строится <u>правосторонняя</u> критическая область, т.е. критическая область расположена справа от критической точки:



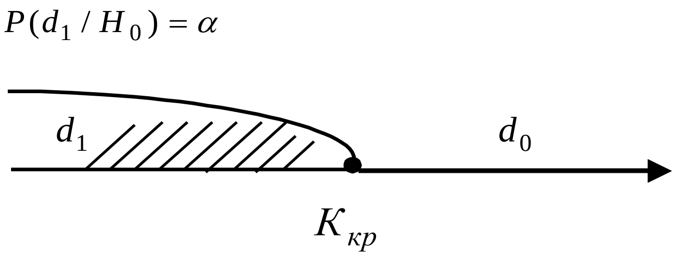
Правосторонняя критическая область строится и при проверке гипотез о виде закона распределения.

Если $H_0 : \theta = \theta_0$, а $H_1 : \theta \neq \theta_0$, строится двусторонняя критическая область:



Заметим, что в случае двусторонней критической области, область принятия нулевой гипотезы H_0 совпадает с интервальной оценкой или доверительным интервалом для параметра θ, который накрывает неизвестное значение этого параметра с вероятностью $1-\alpha$. Если предполагаемое в основной гипотезе числовое значение неизвестного параметра θ_0 попадает в интервальную оценку этого параметра (или интервальная оценка содержит гипотетическое значение параметра θ_0), то гипотезу $H_0 : \theta = \theta_0$ принимают; в противном случае ее отклоняют в пользу $H_1 : \theta \neq \theta_0$ при заданном уровне значимости α.

И, наконец, если $H_0 : \theta = \theta_0$, $H_1 : \theta < \theta_0$, критическая область d_1 будет левосторонней:



5. По данным выборки $\{x_1, \ldots, x_n\}$ подсчитывается наблюдаемое или экспериментальное значение критерия K_0 (или $K_{набл}, K_{эксп}$). Если окажется, что вычисленное значение критерия принадлежит области принятия нулевой гипотезы ($K_0 \in d_0$), то H_0 следует принять, т.е. считать ее не противоречащей выборочным данным. В противном случае гипотезу H_0 следует отвергнуть.

Так, например, если в случае правосторонней критической области $K_0 < K_{кр}$, решение выносится в пользу H_0; если $K_0 > K_{кр}$, то решение выносится в пользу H_1; если $K_0 = K_{кр}$, то наступает рандомизация, т.е. решение в пользу H_0 выносится на основе некоего эксперимента со случайным исходом, практически же, меняют уровень значимости α.

3.3. Гипотезы о параметрах нормального распределения

Одной из наиболее часто встречающихся задач является статистическая проверка гипотез о параметрах нормального распределения.

Пусть $\{x_1, \ldots, x_n\}$ - независимая выборка из нормальной генеральной совокупности, т.е. исследуемая случайная величина $X \sim N(a, \sigma)$, где $a = MX$, $\sigma = \sqrt{DX}$. Здесь возможны следующие предположения о значениях неизвестных параметров.

1. Гипотезы о неизвестном математическом ожидании нормальной генеральной совокупности при: а) σ известном; б) σ неизвестном.

2. Гипотезы о неизвестной дисперсии нормального распределения при: а) a известном; б) a неизвестном.

Критерии проверки гипотез о числовых значениях параметров нормального распределения приведены в табл. 2. В этой же таблице приведен критерий проверки гипотезы о значении вероятности успеха в единичном испытании. В качестве критических точек взяты квантили соответствующих распределений. Квантили распределений χ^2, Стьюдента, Фишера приведены в приложениях 4 – 6.

Таблица 2

Критерии значимости для проверки гипотез о параметрах нормального распределения

Проверяемая гипотеза H_0, H_1	Предположения	Статистика критерия	Распределение статистики	Область отклонения H_0
$H_0: a = a_0$, $H_1: a > a_0$	σ^2 известно	$U = \dfrac{\bar{x} - a_0}{\sigma}\sqrt{n}$	$N(0,1)$	$U_0 > U_{1-\alpha}$
$H_0: a = a_0$, $H_1: a \neq a_0$				$\lvert U_0 \rvert > U_{1-\frac{\alpha}{2}}$
$H_0: a = a_0$, $H_1: a < a_0$				$U_0 < -U_{1-\alpha}$
$H_0: a = a_0$, $H_1: a > a_0$	σ^2 неизвестно $\hat{\sigma}^2 = s^2$	$t = \dfrac{\bar{x} - a_0}{s}\sqrt{n}$	$t(n-1)$	$t_0 > t_{1-\alpha}(n-1)$
$H_0: a = a_0$, $H_1: a \neq a_0$				$\lvert t_0 \rvert > t_{1-\frac{\alpha}{2}}(n-1)$

$H_0: a = a_0$, $H_1: a < a_0$				$t_0 < -t_{1-\alpha}(n-1)$
$H_0: \sigma^2 = \sigma_0^2$, $H_1: \sigma^2 > \sigma_0^2$	a известно	$\chi^2 = \dfrac{ns^2}{\sigma_0^2}$	$\chi^2(n)$	$\chi_0^2 > \chi_{1-\alpha}^2(n)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$				$\chi_0^2 > \chi_{1-\frac{\alpha}{2}}^2(n)$ $\chi_0^2 < \chi_{\frac{\alpha}{2}}^2(n)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$				$\chi_0^2 < \chi_{\alpha}^2(n)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 > \sigma_0^2$	a неизв-но $\hat{a} = \bar{x}$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	$\chi^2(n-1)$	$\chi_0^2 > \chi_{1-\alpha}^2(n-1)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 \neq \sigma_0^2$				$\chi_0^2 > \chi_{1-\frac{\alpha}{2}}^2(n-1)$ $\chi_0^2 < \chi_{\frac{\alpha}{2}}^2(n-1)$
$H_0: \sigma^2 = \sigma_0^2$ $H_1: \sigma^2 < \sigma_0^2$				$\chi_0^2 < \chi_{\alpha}^2(n-1)$
$H_0: p = p_0$, $H_1: p > p_0$	$n > 50$, $np_0 > 5$, $n(1-p_0)>5$	$Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}}$ $\hat{p} = \dfrac{\mu}{n}$, $q_0 = 1 - p_0$	$N(0,1)$	$Z_0 > Z_{1-\alpha}$
$H_0: p = p_0$, $H_1: p \neq p_0$				$\vert Z_0 \vert > Z_{1-\frac{\alpha}{2}}$
$H_0: p = p_0$, $H_1: p < p_0$				$Z_0 < -Z_{1-\alpha}$
Проверяемая гипотеза H_0, H_1	Предположения	Статистика критерия	Распределение статистики	Область отклонения H_0

<u>Пример 11.</u> Крупная торговая фирма желает открыть в новом районе города филиал. Известно, что фирма будет работать прибыльно, если еженедельный средний доход жителей города превышает 400 д.е. Известно также,

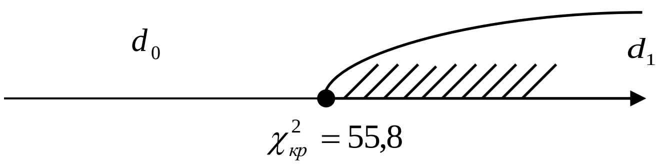
что дисперсия дохода $\sigma^2 = 400$. Определить правило принятия решения, с помощью которого, основываясь на выборке $n = 100$ и уровне значимости $\alpha = 0{,}05$, можно установить, что филиал будет работать прибыльно.

Решение. Определим правило принятия решения, основываясь на статистической проверке гипотез. Фирма не откроет филиал, если средний доход жителей не превысит 400 д.е. Будем считать, что доход является нормально распределенной случайной величиной и $H_0: a = 400$, а $H_1: a > 400$. Значение σ^2 дисперсии дохода известно, в этом случае H_1 принимают, если $U_0 = \dfrac{\bar{x} - a_0}{\sigma}\sqrt{n} > U_{1-\alpha}$. По условию $a = 400, \sigma = 20,$ $\sqrt{n} = 10,$ $U_{1-0,05} = U_{0,95} = 1{,}65$ (этот квантиль уровня 0,05 стандартного нормального закона может быть найден из таблиц значений функции Лапласа (см. приложение 3) как такой ее аргумент, при котором она равна $0{,}5 - \alpha$, т.е. $U_{0,95} : \Phi(U) = 0{,}45$). Поэтому H_1 принимают, и, следовательно, филиал открывают, если недельный среднедушевой доход 100 жителей будет $\bar{x} > 400 + 2 \cdot 1{,}65 = 403{,}3.$

Пример 12. Партия изделий принимается, если дисперсия контролируемого размера не превышает 0,2. По выборке $n = 40$ изделий вычислена $s^2 = 0{,}25$. Можно ли принять партию при $\alpha = 0{,}05$?

Решение. Следуя общей логической схеме проверки гипотез, имеем

1. $H_0 : \sigma^2 = 0{,}2$ (или $H_0 : \sigma^2 \le 0{,}2$)
   $H_1 : \sigma^2 > 0{,}2$;
2. $\alpha = 0{,}05$;
3. Статистика критерия $\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1)$;
4. Критическую точку $\chi^2_{кр}$ найдем из таблицы квантилей распределения χ^2 (см. приложение 4), как квантиль уровня $1 - \alpha = 0{,}95$ при числе степеней свободы, равном $n - 1 = 39: \chi^2_{кр} = \chi^2_{0,95}(39) = 55{,}8.$ Критическая область в нашем случае правосторонняя:



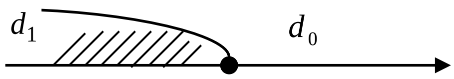
5. По результатам наблюдений определим экспериментальное значение критерия $\chi_0^2 = \dfrac{(40-1)0{,}25}{0{,}2} = 48{,}75.$ Так как $\chi_0^2 < \chi^2_{кр}$, т.е. наблюдаемое зна-

чение критерия принадлежит области принятия нулевой гипотезы, то H_0 следует принять, т.е. считать, что различие между гипотетическим значением дисперсии, равным 0,2 и его оценкой 0,25 статистически незначимо или случайно, и поэтому всю партию изделий можно принять, допуская при этом ошибку первого рода с вероятностью 0,05.

<u>Пример 13.</u> Торговец утверждает, что он получает заказы в среднем по крайней мере от 30% предполагаемых клиентов. Можно ли при 5%-ом уровне значимости считать это утверждение неверным, если торговец получил заказы от 20 из 100 случайно отобранных потенциальных клиентов.

<u>Решение.</u> Так как $\hat{p} = \dfrac{20}{100} = 0{,}2$, то задача формулируется следующим образом:

1. $H_0 : p = 0{,}3, \quad H_1 : p < 0{,}3.$
2. $\alpha = 0{,}05$.
3. Статистика критерия $Z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0 q_0}{n}}} \sim N(0{,}1).$
4. Критическая область в данном случае левосторонняя $-U_{0,95} = -1{,}65$:



$-U_{0,95} = -1{,}65$

6. Экспериментальное значение критерия $Z_0 = \dfrac{0{,}2 - 0{,}3}{\sqrt{\dfrac{0{,}3 \cdot 0{,}7}{100}}} = -2{,}18.$

Так как Z_0 принадлежит критической области d_1, то гипотезу H_0 отклоняем, т.е. с утверждением торговца нельзя согласиться.

<u>Пример 14.</u> По паспортным данным автомобильного двигателя расход топлива на 100 км пробега составляет 10 л. В результате изменения конструкции двигателя ожидается, что расход топлива уменьшится. Для проверки проводятся испытания 25 случайно отобранных автомобилей с модернизированным двигателем, причем выборочное среднее расходов топлива на 100 км пробега по результатам испытаний составило $\bar{x} = 9{,}3$л. Предположим, что выборка расходов топлива получена из нормального распределения генеральной совокупности со средним a и дисперсией $\sigma^2 = 4$. Используя критерий значимости, проверить гипотезу о том, что изменение конструкции двигателя не повлияло на расход топлива.

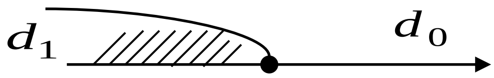
Решение. Проверяется гипотеза о среднем нормально распределенной генеральной совокупности. Проверку гипотезы проведем по этапам.

1. $H_0 : a = 10, \quad H_1 : a < 10.$

2. $\alpha = 0{,}05$.

3. Статистика критерия $U_0 = \dfrac{\bar{x} - a_0}{\sigma}\sqrt{n} \sim N(0{,}1).$

4. Критическую точку найдем из таблиц квантилей стандартного нормального закона $-U_{1-\alpha} = -U_{0,95} = -1{,}65$. Критическая область в нашем случае левосторонняя:



$-U_{0,95} = -1{,}65$

5. Экспериментальное значение критерия $U_0 = \dfrac{9{,}3 - 10}{2}\sqrt{25} = -1{,}75$, так как $U_0 \in d_1$, то гипотеза H_0 отклоняется, т.е. следует считать, что изменение конструкции двигателя привело к уменьшению расхода топлива.

Задачи

44. Техническая норма предусматривает в среднем 40 с. на выполнение определенной технологической операции на конвейере по производству часов. От работающих на этой операции поступили жалобы, что они в действительности затрачивают на нее больше времени. Для проверки этой жалобы были произведены хронометрические измерения времени выполнения этой технологической операции у 16 работниц, занятых на ней, и получено среднее время операции 42 с. Можно ли по имеющимся хронометрическим данным на уровне значимости $\alpha = 0{,}01$ отклонить гипотезу о том, что среднее время выполнения этой операции соответствует норме, если $\sigma = 3{,}5$ с?

45. Компания, производящая средства для потери веса утверждает, что прием таблеток в сочетании со специальной диетой позволяет сбросить в среднем в неделю 400 г веса. Случайным образом отобраны 25 человек, использующих эту терапию, и обнаружено, что в среднем еженедельная потеря в весе составила 430 г со средним квадратическим отклонением 110 г. Проверить гипотезу о том, что средняя потеря веса составляет 400 г ($\alpha = 0{,}05$).

46. Компания, выпускающая в продажу новый сорт растворимого кофе, провела проверку вкусов покупателей по случайной выборке из 400 человек и выяснила, что 220 из них предпочли новый сорт всем остальным. Проверить на уровне значимости $\alpha = 0{,}01$ гипотезу о том, что по крайней мере 52% потребителей предпочтут новый сорт кофе.

47. Страховая компания изучает вероятность ДТП для подростков, имеющих мотоциклы. За прошедший год проведена случайная выборка 2000 страховых полисов подростков-мотоциклистов и выявлено, что 15 из них попадали в ДТП и предъявили компании требование о компенсации за ущерб. Может ли аналитик компании отклонить гипотезу о том, что менее 1% всех подростков-мотоциклистов, имеющих страховые полисы, попали в ДТП в прошлом году ($\alpha = 0{,}05$)?

48. Компания по производству безалкогольных напитков предполагает выпустить на рынок новую модификацию популярного напитка, в котором сахар заменен сукразитом. Компания хотела бы быть уверенной в том, что не менее 70% ее потребителей предпочтут новую модификацию напитка. Новый напиток был предложен на пробу 2000 человек, и 1422 из них сказали, что он вкуснее старого. Может ли компания отклонить предположение о том, что только 70% всех ее потребителей предпочтут новую модификацию напитка старой ($\alpha = 0{,}05$)?

49. Производитель нового типа аспирина утверждает, что он снимает головную боль за 30 мин. Случайная выборка из 100 человек, страдающих головными болями, показала, что новый тип аспирина снимает головную боль за 28,6 мин. при среднем квадратическом отклонении 4,2 мин. Проверить на уровне значимости $\alpha = 0{,}05$ справедливость утверждения производителей аспирина о том, что это лекарство излечивает головную боль за 30 мин.

50. Компания, занимающаяся консультированием в области инвестиций, заявляет, что среднегодовой процент по акциям определенной отрасли промышленности составляет 11,5%. Инвестор, желая проверить истинность этого утверждения, на основе случайной выборки 50 акций выявил, что среднегодовой процент по ним составил 10,8% с исправленным средним квадратическим отклонением $s = 3{,}4\%$. На основе имеющейся информации определить, имеет ли инвестор достаточно оснований, чтобы отвергнуть заявление компании ($\alpha = 0{,}05$)?

51. Производитель некоторого вида продукции утверждает, что 95% выпускаемой продукции не имеет дефекта. Случайная выборка 100 изделий показала, что только 92 из них свободны от дефекта. Проверить справедливость утверждения производителя на уровне значимости $\alpha = 0{,}05$.

52. Считается, что завод, производящий за неделю 1000 телевизоров, работает удовлетворительно, если в среднем частость бракованных телевизоров не превышает 3%. Допустим, что в течение некоторой недели было забраковано 38 телевизоров. Необходимо ли директору завода провести более пол-

ную проверку качества телевизионной производственной линии или же следует отнести высокий процент дефектной продукции этой недели за счет случайных изменений в условиях производства ($\alpha = 0{,}05$)?

53. По утверждению фирмы, средний размер дебиторского счета 187,5 т.р. Ревизор составляет случайную выборку из 50 счетов и обнаруживает, что среднее арифметическое выборки равно 175 т.р. при среднем квадратическом отклонении 35 т.р. Может ли оказаться в действительности правильным объявленный размер дебиторского счета ($\alpha = 0{,}05$)?

54. Фирма, производящая электрические лампочки, утверждает, что среднее время безотказной работы лампочек $W = 120\ Bm$ равно 800 часов со $\sigma = 120$ часов. Из некоторой партии таких лампочек производится выборка 25 лампочек, для которой выборочное среднее времени работы лампочек оказалось 750 часов. Можно ли на основании этого сказать, что исследуемая выборка лампочек не удовлетворяет гарантии ($\alpha = 0{,}05$)?

55. Высота отдельных ростков рассады распределена нормально со средней a = 53 см и дисперсией $\sigma^2 = 12\,\text{см}^2$. В прошлом году в ящик, в котором были высажены n = 15 таких растений, была внесена по ошибке двойная норма удобрения. Средняя высота рассады в этом ящике достигла $\bar{x} = 55$ см. Есть ли какое-либо основание предполагать, что повышенное внесение удобрений дало положительный эффект?

3.4. Гипотезы о равенстве средних и дисперсий двух нормальных распределений

Задачи, связанные с проверкой гипотез о равенстве средних и дисперсий двух нормальных генеральных совокупностей, возникают при сравнении способов управления производством, различных технологических процессов или методов обработки по определенным измеряемым признакам (точности, производительности и т.д.). Например, предположим, что компания производит определенный элемент на двух автономных производственных линиях – А и В. Характеристики обеих линий одинаковые. Как определить, одинакова ли вариация продукции на этих линиях? Ответ на этот вопрос можно получить, сравнив дисперсии случайных выборок, взятых из продукций первой и второй линий, используя соответствующую процедуру проверки гипотез. Так же можно сравнить риск двух различных инвестиционных портфелей. Сравнение дисперсий фактической прибыли, полученной в прошлые годы, даст возможность принять решение.

Пусть имеются две независимые выборки $\{x_1, x_2, \ldots, x_{n_1}\}$ и $\{y_1, y_2, \ldots, y_{n_2}\}$, извлеченные из нормальных генеральных совокупностей, т.е. исследуемые случайные признаки $X \sim N(a_1, \sigma_1)$, $Y \sim N(a_2, \sigma_2)$. Здесь возможны следующие предположения о значениях неизвестных параметров.

1. Гипотеза о равенстве средних при известных дисперсиях $H_0 : a_1 = a_2$, если σ_1, σ_2 известны.

2. Гипотеза о равенстве средних при неизвестных дисперсиях $H_0 : a_1 = a_2$, если σ_1, σ_2 неизвестны.

3. Гипотеза о равенстве дисперсий при известных средних $H_0 : \sigma_1^2 = \sigma_2^2$, если a_1, a_2 известны.

4. Гипотеза о равенстве дисперсий при неизвестных средних $H_0 : \sigma_1^2 = \sigma_2^2$, если a_1, a_2 неизвестны.

Критерии проверки таких основных гипотез при различных конкурирующих гипотезах приведены в табл. 3.

Замечание 1. Если гипотезу $H_0 : a_1 = a_2$ принимают, то говорят, что различие выборочных средних $\bar{x}$ и $\bar{y}$ статистически не значимо и оценка общего математического ожидания такова: $(\bar{x}n_1 + \bar{y}n_2) \big/ (n_1 + n_2)$.

Замечание 2. При проверке гипотезы $H_0 : a_1 = a_2$ при неизвестных дисперсиях, сначала проверяют гипотезу о равенстве дисперсий $H_0 : \sigma_1^2 = \sigma_2^2$, если этот факт заранее не известен. В случае приемлемости гипотезы $H_0 : \sigma_1^2 = \sigma_2^2$, приступают к проверке гипотезы $H_0 : a_1 = a_2$.

Замечание 3. Если гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ принимают, то говорят, что различие оценок дисперсий s_1^2 и s_2^2 статистически незначимо и оценка общей дисперсии такова: $\left[s_1^2(n_1 - 1) + s_2^2(n_2 - 1)\right] \big/ (n_1 + n_2 - 2)$.

Таблица 3

Критерии проверки гипотез о средних и дисперсиях двух нормальных распределений

Проверяемая гипотеза H_0, H_1	Предположения	Статистика критерия	Распределение статистики	Область отклонения H_0
$H_0 : a_1 = a_2$, $H_1 : a_1 > a_2$				$U_0 > U_{1-\alpha}$

$H_0: a_1 = a_2$, $H_1: a_1 \neq a_2$	σ_1^2, σ_2^2 известны	$U = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$N(0,1)$	$\lvert U_0 \rvert > U_{1-\frac{\alpha}{2}}$
$H_0: a_1 = a_2$, $H_1: a_1 < a_2$				$U_0 < -U_{1-\alpha}$
$H_0: a_1 = a_2$, $H_1: a_1 > a_2$	σ_1^2, σ_2^2 неизвест-ны, но равны	$t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$	$t(n_1 + n_2 - 2)$	$t_0 > t_{1-\alpha}(n_1 + n_2 - 2)$
$H_0: a_1 = a_2$, $H_1: a_1 \neq a_2$				$\lvert t_0 \rvert > t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$
$H_0: a_1 = a_2$, $H_1: a_1 < a_2$				$t_0 < -t_{1-\alpha}(n_1 + n_2 - 2)$
$H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 > \sigma_2^2$	a_1, a_2 известны	$F = \dfrac{s_1^2}{s_2^2}$, $s_1^2 > s_2^2$	$F(n_1, n_2)$	$F_0 > F_{1-\alpha}(n_1, n_2)$
$H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$				$F_0 > F_{1-\frac{\alpha}{2}}(n_1, n_2)$ $F_0 < F_{\frac{\alpha}{2}}(n_1, n_2)$
$H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 < \sigma_2^2$		$F = \dfrac{s_1^2}{s_2^2}$, $s_2^2 > s_1^2$	$F(n_1, n_2)$	$F_0 > F_{1-\alpha}(n_1, n_2)$
$H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 > \sigma_2^2$	a_1, a_2 неиз-вестны $\hat{a}_1 = \bar{x}_1$ $\hat{a}_2 = \bar{x}_2$	$F = \dfrac{s_1^2}{s_2^2}$, $s_1^2 > s_2^2$	$F(n_1 - 1, n_2 - 1)$	$F_0 > F_{1-\alpha}(n_1 - 1, n_2 - 1)$
$H_0: \sigma_1^2 = \sigma_2^2$, $H_1: \sigma_1^2 \neq \sigma_2^2$				$F_0 > F_{1-\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$ $F_0 < F_{\frac{\alpha}{2}}(n_1 - 1, n_2 - 1)$

$H_0:\sigma_1^2=\sigma_2^2$, $H_1:\sigma_1^2<\sigma_2^2$		$F=\frac{s_1^2}{s_2^2}$, $s_2^2>s_1^2$	$F(n_1-1,$ $n_2-1)$	$F_0>F_{1-\alpha}(n_1-1,$ $n_2-1)$
Проверяемая гипотеза H_0, H_1	Предпо-ложения	Статистика критерия	Распреде-ление ста-тистики	Область откло-нения H_0

Пример 15. Биржевой маклер исследует две инвестиции А и В от имени клиента. Инвестиция А предполагается на срок 10 лет с ожидаемой ежегодной прибылью в течение этого периода 17,8%. Инвестиция В рассчитана на срок 8 лет также с ожидаемой годовой прибылью 17,8%. Дисперсии ежегодных прибылей от двух инвестиций составляют $(3{,}21\%)^2$ и $(7{,}14\%)^2$. Есть ли какое-либо основание считать, что риски инвестиций А и В не равны? Предполагается, что ежегодные прибыли от инвестиций нормально распределены.

Решение. Дисперсии ежегодных прибылей могут быть использованы для определения риска. Поэтому для того, чтобы ответить на вопрос задачи, мы должны проверить статистическую гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей. Следуя общей схеме проверки гипотез, имеем:

1. $H_0:\sigma_A^2=\sigma_B^2,\quad H_1:\sigma_A^2\neq\sigma_B^2.$

2. $\alpha=0{,}05.$

3. Статистика критерия $F=\dfrac{s_б^2}{s_м^2}\sim F(n_1-1,n_2-1)$, где $s_б^2$ –большая оценка дисперсии, $s_м^2$ –меньшая оценка теоретической или генеральной дисперсии; n_1,n_2 –объемы выборок, по которым найдены соответствующие оценки.

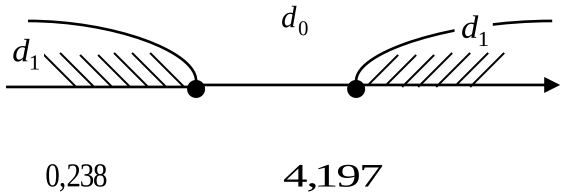
4. Так как $H_1:\sigma_A^2\neq\sigma_B^2$, критическая область будет двусторонняя. Для того чтобы определить значения критических точек из таблицы квантилей F-распределения (см. приложение 6), найдем несмещенные оценки теоретических дисперсий:

$$\hat{\sigma}_A^2=s_A^2=\frac{n_A}{n_A-1}D_{BA}=\frac{10}{9}(3{,}21)^2=11{,}449=s_м^2,$$

$$\hat{\sigma}_B^2=s_B^2=\frac{n_B}{n_B-1}D_{BB}=\frac{8}{7}(7{,}14)^2=58{,}2624=s_б^2.$$

Таким образом, числа степеней свободы для F-распределения будут 7 и 9. Имеем $F_{0,975}(7{,}9)=4{,}197=F_{кр.прав}$, левая точка определится как $F_{кр.лев}=\dfrac{1}{F_{кр.прав}}=\dfrac{1}{4{,}197}=0{,}238.$

Критическая область:



5. Экспериментальное значение критерия

$$F_0 = \frac{s_б^2}{s_м^2} = \frac{58{,}2624}{11{,}449} = 5{,}09.$$

Так как 5,09>4,197, т.е. $F_0 \in d_1$, гипотезу H_0 следует отвергнуть на 5%-м уровне значимости. Следовательно, у нас есть основания предполагать, что риски (определенные дисперсиями ежегодных прибылей) двух инвестиций не равны.

Пример 16. Расход сырья на одно изделие случаен. Результаты наблюдений таковы:

Расход сырья	Старая технология			Новая технология			
	304	307	308	303	304	306	308
Число изделий	1	4	4	2	6	4	1

Предположив, что расход сырья как при старой, так и при новой технологии имеет нормальное распределение, выяснить, влияет ли технология на средний расход сырья на одно изделие. Принять $\alpha = 0{,}05$.

Решение. Для того чтобы ответить на вопрос задачи, необходимо проверить гипотезу о равенстве средних двух нормальных генеральных совокупностей, дисперсии которых не известны и неизвестно, равны ли они. Поэтому, прежде чем сравнивать генеральные средние, проверим гипотезу $H_0' : \sigma_1^2 = \sigma_2^2$. Найдем по данным выборки несмещенные оценки средних и дисперсий:

$$\bar{x} = \frac{1}{9}(304 + 307 \cdot 4 + 308 \cdot 4) = 307{,}11,$$

$$\bar{y} = \frac{1}{13}(303 \cdot 2 + 304 \cdot 6 + 306 \cdot 4 + 308) = 304{,}77,$$

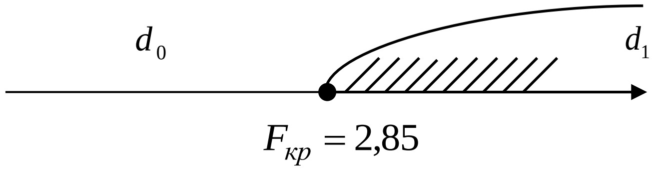
$$s_1^2 = 2{,}378, \quad s_2^2 = 1{,}685.$$

Следуя общей схеме проверки гипотез, имеем

1. $H_0' : \sigma_1^2 = \sigma_2^2, \quad H_1' : \sigma_1^2 > \sigma_2^2.$
2. $\alpha = 0{,}05.$

3. $F = \dfrac{s_1^2}{s_2^2} \sim F(n_1 - 1, n_2 - 1)$, где $n_1 = 9$, $n_2 = 13$.

4. Из таблицы квантилей F-распределения (см. приложение 6), найдем критическую точку $F_{0,95}(8,12) = 2,85 = F_{кр}$. Критическая область правосторонняя:



5. Наблюдаемое значение критерия равно $F_0 = \dfrac{2,378}{1,685} = 1,41$.

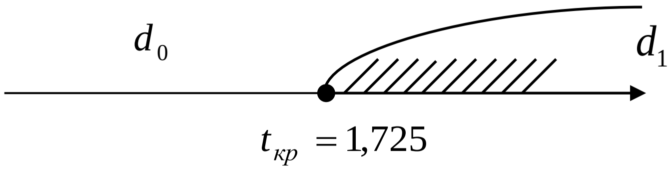
Так как 1,41<2,85, т.е. $F_0 \in d_0$, гипотезу о равенстве дисперсий принимаем.

Теперь проверим гипотезу:

1. $H_0 : a_1 = a_2$, $\quad H_1 : a_1 > a_2$.
2. $\alpha = 0,05$.
3. $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$, где $n_1 = 9$, $n_2 = 13$,

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

4. Из таблицы квантилей t-распределения (см. приложение 5) найдем критическую точку $t_{кр} = t_{0,95}(20) = 1,725$. Критическая область:



5. Для расчета наблюдаемого значения вычислим сначала $s^2 = (8 \cdot 2,378 + 12 \cdot 1,685)/20 = 1,9622$. Тогда

$$t_0 = \frac{307,11 - 304,77}{\sqrt{1,9622\left(\dfrac{1}{9} + \dfrac{1}{13}\right)}} = 3,852.$$

Так как 3,852>1,725, т.е. $t_0 \in d_1$, считаем, что применение новой технологии снижает затраты сырья на одно изделие.

Задачи

56. В двух фирмах, выпускающих детское питание, производилась оценка качества продукции. В фирме *A*, где проверялось 30 единиц продукции, средняя сумма баллов оказалась равной 52. Во второй фирме В проверялось 36 единиц продукции, и их средняя сумма баллов оказалась равной 47. Среднее квадратическое отклонение суммы баллов, вычисленное для нескольких тысяч единиц продукции, $\sigma = 12$. Определить, лучшее ли питание выпускается фирмой *A*, чем фирмой В.

57. Предполагается, что применение нового типа резца сократит время обработки некоторой детали. Хронометраж времени обработки 9 деталей, обработанных старым типом резцов, дал следующие результаты: $\bar{x} = 57$ мин., $s_1^2 = 186{,}2\,\text{м}^2$. Среднее время обработки 15 деталей, обработанных новым типом резцов, $\bar{y} = 52$ мин., $s_2^2 = 166{,}4\,\text{м}^2$. На уровне значимости $\alpha = 0{,}05$ ответить на вопрос, позволило ли использование нового типа резцов сократить время обработки детали?

58. Покрышки автомобиля были исследованы на износ после 10000 км пути. Каждая покрышка сделана наполовину из резины сорта А, наполовину – из резины сорта В. Все колеса вращаются с одинаковой скоростью, поэтому можно считать, что они находятся в одинаковых условиях. Результаты исследования (в баллах) приведены в таблице

Номер покрышки	1	2	3	4
Износ резины сорта А	32	40	36	35
Износ резины сорта В	25	28	27	26

Проверить гипотезу о том, что резина сорта В изнашивается больше, чем резина сорта А.

59. В магазине отобрали случайным образом по 40 банок «Солянка» производства двух заводов. Каждая из банок была оценена с помощью специальных показателей. Результаты оценки приведены в таблице

	№ 1	№ 2
Средний балл	71	76
Стандартное отклонение	5	6

Проверить гипотезу на уровне $\alpha = 0{,}05$ о том, что консервы завода №2 лучшего качества, чем завода №1.

60. Студенты экономического и технического вузов сдавали экзамены по математике. В экономическом университете, где экзаменовалось 30 студентов, средняя оценка оказалась 4,52. В техническом вузе сдавали экзамен 36 студентов, их средняя оценка оказалась равной 4,47. Среднее квадратическое отклонение оценок равно 0,12. Лучше ли подготовлены студенты экономического университета, чем технического?

61. Компания по производству сахарного песка имеет две производственные линии для наполнения пакетов сахарным песком по 1 кг (a_1 и a_2). Используя данные, собранные в течение многих лет, управляющий оценивает стандартное отклонение веса пакетов, поставляемых с первой линии, в 0,02 кг (σ_1) и со второй линии – в 0,04 кг (σ_2). Из продукции первой линии была взята случайная выборка объема $n_1 = 10$ пакетов и найден средний вес содержимого в них сахарного песка $\bar{x}_1 = 1{,}018$ кг. Подобная выборка $n_2 = 12$ пакетов была взята со второй линии и найден средний вес $\bar{x}_2 = 0{,}989$ кг. Имеется ли основание предположить, что две производственные линии развешивают сахарный песок в пакеты с различным средним весом?

62. Внешний аудитор проверяет систему учета малой компании. Первая проверка 100 сделок показала, что 56 сделок ошибочны. Компания предоставила месяц для исправления системы. Когда аудитор провел вторую проверку, он обнаружил, что 28 сделок ошибочны из 78 проведенных сделок. Есть ли какое-либо основание предполагать, что доля ошибок между проверками уменьшилась?

3.5. Гипотезы о виде закона распределения. Критерий согласия χ^2

Рассмотренные в п.3.3. и 3.4. методы проверки статистических гипотез предполагали известным закон распределения и касались лишь значений параметров этого закона. Однако в ряде случаев сама форма закона распределения является гипотетической и нуждается в проверке. Здесь речь пойдет о проверке основной гипотезы H_0 о том, что данная случайная величина X подчинена закону распределения $F_0(x)$. Конкурирующая альтернатива H_1 заключается здесь просто в отклонении H_0. Критерии проверки таких гипотез, называемые обычно критериями согласия, основаны на выборе определенной меры расхождения между теоретическим (или гипотетическим) и эмпирическим распределениями. Одним из наиболее распространенных является критерий Пирсона.

В <u>критерии Пирсона (критерий χ^2)</u> за меру расхождения статистического и теоретического законов распределения принимается величина χ^2, выборочное значение которой определяется формулой

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i},$$

где k – число различных вариант (число интервалов группирования); n – объем выборки. Если X – дискретная случайная величина, то p_i – вероятность реализации значения x_i, вычисленная в предположении, что выдвигаемая гипотеза верна, т.е. $p_i = P\{X = x_i / H_0\}$. Если X – непрерывная случайная величина, то p_i – вероятность попадания в i – й интервал, $p_i = P\{x_i < X < x_{i+1} / H_0\}$, $i = 1, \ldots, k$. Очевидно, что в обоих случаях $\sum_{i=1}^{k} p_i = 1$.

При $n \to \infty$ закон распределения статистики χ^2 независимо от вида закона распределения X стремится к закону $\chi^2(q), q = k - r - 1$, где r – число параметров теоретического распределения. Так, если выдвигается гипотеза о принадлежности случайной величины к нормальному или равномерному закону, то $q = k - 3$; если гипотетический закон – распределение Пуассона, то $q = k - 2$.

Процедура применения критерия χ^2 для проверки гипотезы H_0 о том, что исследуемая случайная величина X имеет закон распределения $F_0(x)$, состоит из следующих этапов.

1. По выборке $\{x_1, \ldots, x_n\}$ наблюдений случайной величины X найти оценки неизвестных параметров предполагаемого закона распределения $F_0(x)$.
2. Получить эмпирическое распределение случайной величины в виде точечного или интервального вариационных рядов.
3. Определить теоретические вероятности p_i в предположении, что выдвигаемая гипотеза верна.
4. Вычислить наблюдаемое или экспериментальное значение статистики критерия χ_0^2.
5. Принять статистическое решение: гипотеза H_0 не противоречит выборке наблюдений при заданном уровне значимости α, если $\chi_0^2 < \chi_{кр}^2$, где критическая точка $\chi_{кр}^2 = \chi_{1-\alpha}^2$ – квантиль уровня $1-\alpha$ распределения χ^2 с числом степеней свободы $k - r - 1$ (см. приложение 4). Если же $\chi_0^2 > \chi_{кр}^2$, то гипотеза H_0 отклоняется.

Рассмотрим этапы использования критерия χ^2 на примере проверки гипотезы о нормальном распределении.

Пусть $H_0 : X \sim N(a,\sigma)$, где параметры $a = MX$ и $\sigma = \sqrt{DX}$ неизвестны. По независимой выборке $\{x_1,\dots,x_n\}$ наилучшими оценками этих параметров будут соответственно $\hat{a} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ и $\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

Если выборка представлена в виде последовательности k интервалов шириной h, то несмещенной оценкой математического ожидания и оценкой среднего квадратического отклонения будут $\hat{a} = \bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^*$, $\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i^* - \bar{x})^2 - \frac{h^2}{12}}$, здесь $x_i^* = \frac{x_i + x_{i+1}}{2}$ – середина i – го интервала, n_i – соответствующая частота, $\sum_{i=1}^{k} n_i = n$.

Статистика критерия $\chi^2 = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$ распределена по закону χ^2 с $k-3$ степенями свободы. Теоретические вероятности p_i в предположении, что гипотеза H_0 верна, вычисляются по формуле

$$p_i = P\{x_i < X < x_{i+1}\} = \Phi\left(\frac{x_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{x_i - \bar{x}}{s}\right),$$

где $\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}\,dt$ – функция Лапласа (см. приложение 3).

Наблюдаемое значение χ_0^2 сравнивается с критической точкой $\chi_{кр}^2 = \chi_{1-\alpha}^2(k-3)$. Если $\chi_0^2 < \chi_{кр}^2$, то при заданном уровне значимости α гипотезу H_0 принимают.

Пример 17. По данным примера 1 выяснить, можно ли на уровне значимости $\alpha = 0{,}05$ считать нормальным распределение производительности труда.

Решение. На принадлежность к нормальной генеральной совокупности исследуемой выборки объема $n = 100$ указывали элементы первичной статистической обработки данных. Убедимся в этом, используя критерий согласия χ^2. Итак, имеем:

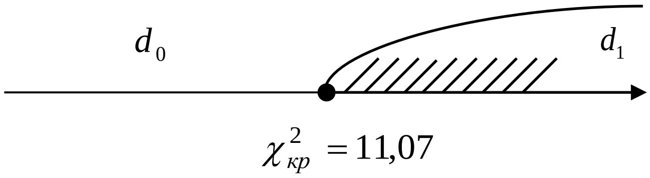
1. $H_0 : X \sim N(a,\sigma^2)$, где $\hat{a} = \bar{x} = 14{,}978$,

$\hat{\sigma} = s = \sqrt{\frac{n}{n-1}D_в} = \sqrt{\frac{100}{99}0{,}9191} = \sqrt{0{,}928} = 0{,}963$.

2. $\alpha = 0{,}05$.

3. $\chi^2 = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(8-3)$ (здесь число интервалов группирования $k = 8$).

4. Из таблиц квантилей распределения χ^2 найдем критическую точку $\chi^2_{кр} = \chi^2_{0,95}(5) = 11{,}07$ (см. приложение 4). Критическая область правосторонняя



5. Для расчета наблюдаемого значения критерия χ^2_0 составим две вспомогательные таблицы

Таблица 4

Расчет $np_i = n\left[\Phi\left(\frac{C_{i+1} - \bar{x}}{s}\right) - \Phi\left(\frac{C_i - \bar{x}}{s}\right)\right]$

i	C_i	C_{i+1}	$z_i = \frac{C_i - \bar{x}}{s}$	$z_{i+1} = \frac{C_{i+1} - \bar{x}}{s}$	$\Phi(z_{i+1})$	$\Phi(z_i)$	p_i	np_i
1	12,5	13,1	-2,52	-1,95	-0,4744	-0,4949	0,0205	2,05
2	13,1	13,7	-1,95	-1,33	-0,4082	-0,4744	0,0662	6,62
3	13,7	14,3	-1,33	-0,70	-0,2580	-0,4082	0,1502	15,02
4	14,3	14,9	-0,70	-0,08	-0,0319	-0,2580	0,2261	22,61
5	14,9	15,5	-0,08	0,54	0,2054	-0,0319	0,2373	23,73
6	15,5	16,1	0,54	1,16	0,3770	0,2054	0,1716	17,16
7	16,1	16,7	1,16	1,79	0,4633	0,3770	0,0863	8,63
8	16,7	17,3	1,79	2,41	0,4920	0,4633	0,0287	2,87
$\sum$							0,9869 ≈ 1	98,69 ≈ 100

Таблица 5

Расчет $\chi^2_0 = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i}$

i	n_i	np_i	$(n_i - np_i)^2$	$(n_i - np_i)^2 / np_i$
1	2	2,05	0,0025	0,0012
2	8	6,62	2,9044	0,4387
3	15	15,02	0,0004	0,0000

4	20	22,61	6,8121	0,3013
5	26	23,73	5,1529	0,2171
6	17	17,16	0,0256	0,0015
7	8	8,63	0,3969	0,0460
8	4	2,87	1,2769	0,4449
$\sum$	100	≈ 100		1,4507

Сравниваем наблюдаемое значение критерия $\chi_0^2 = 1{,}4507$ с критической точкой $\chi_{кр}^2 = 11{,}07$, так как 1,4507 < 11,07, т.е. $\chi_0^2 \in d_0$, гипотезу о нормальном распределении производительности труда принимаем.

Если эмпирическое распределение задано в виде точечного ряда, то расчет теоретических вероятностей нужно производить по формуле $p_i = \frac{l}{\sigma_B}\varphi(u_i)$, где $l -$разность между соседними вариантами, $u_i = \frac{x_i - \overline{x}}{\sigma_B}$, $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$ (см. приложение 2).

Пример 18. Отдел технического контроля проверил n = 200 партий одинаковых изделий и получил следующее эмпирическое распределение:

x_i	0	1	2	3	4
n_i	116	56	22	4	2

$\sum_{i=1}^{5} n_i = 200$, где x_i - число нестандартных изделий в одной партии, n_i - число партий, содержащих x_i нестандартных изделий.

Требуется, при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X – число нестандартных изделий – распределена по закону Пуассона.

Решение. Неизвестным параметром распределения Пуассона является $\lambda = MX$. Наилучшей его оценкой по данному эмпирическому распределению будет выборочное среднее $\widehat{\lambda} = \overline{x}$ (см. решение примера 3). Найдем

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{5} n_i x_i = \frac{1}{200}\left(116\cdot 0 + 56\cdot 1 + 22\cdot 2 + 4\cdot 3 + 2\cdot 4\right) = 0{,}6.$$

Таким образом, для предполагаемого распределения Пуассона вероятности отдельных частных реализаций значений случайной величины X имеют вид:

$$P_n(x_i) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}, \;\Rightarrow\; P_{200}(i) = \frac{(0{,}6)^i}{i!}e^{-0{,}6},$$ где i = 0, 1, 2, 3, 4.

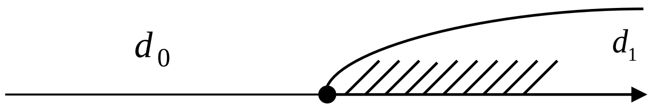
Из таблицы распределения Пуассона (см. приложение 1) для $\lambda = 0{,}6$ выписываем вероятности: $p_0 = P_{200}(0) = 0{,}5488$; $p_1 = P_{200}(1) = 0{,}3293$; $p_2 = P_{200}(2) = 0{,}0988$; $p_3 = P_{200}(3) = 0{,}0198$; $p_4 = P_{200}(4) = 0{,}0030$.

Найдем теоретические частоты по формуле $n_i' = np_i = 200 p_i$.

Подставляя в эту формулу найденные значения вероятностей, получим: $n_0' = 200 \cdot 0{,}5488 = 109{,}76$; $n_1' = 200 \cdot 0{,}3293 = 65{,}86$;
$n_2' = 200 \cdot 0{,}0988 = 19{,}76$; $n_3' = 200 \cdot 0{,}0198 = 3{,}96$; $n_4' = 200 \cdot 0{,}0030 = 0{,}60$.

Представим теперь решение задачи в виде последовательных этапов проверки гипотезы.

1. $H_0 : X \sim \Pi(\lambda)$, где $\hat{\lambda} = 0{,}6$;
2. $\alpha = 0{,}05$;
3. $\chi^2 = \sum_{i=1}^{5} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2(5-2)$;
4. $\chi^2_{кр} = \chi^2_{0,95}(3) = 7{,}81$. Критическая область правосторонняя



$\chi^2_{кр} = 7{,}81$

5. Для расчета наблюдаемого значения критерия χ^2_o составим таблицу:

Таблица 6

Расчет $\chi_0{}^2 = \sum_{i=1}^{5} \frac{(n_i - np_i)^2}{np_i}$

i	n_i	$n_i' = np_i$	$(n_i - n_i')^2$	$(n_i - n_i')^2 / n_i'$
0	116	109,76	38,9376	0,3548
1	56	65,86	97,2196	1,4762
2	22	19,76	5,0176	0,2539
3	4	3,96	0,0016	0,0004
4	2	0,60	1,9600	3,2667
$\sum$				5,3520

Так как $\chi^2_o < \chi^2_{кр}$ (5,352 < 7,81), т.е. $\chi^2_o \in d_o$, то нет оснований отвергнуть гипотезу о распределении числа нестандартных изделий по закону Пуассона.

Задачи

63. Рассматривая распределение размеров мужской обуви, проданной магазином за смену, как выборку из генеральной совокупности, проверить гипотезу о том, что интересующий нас признак распределен в генеральной совокупности по нормальному закону ($\alpha = 0{,}01$). Данные о количестве пар проданной обуви представлены в таблице

Размер обуви, x_i	36	37	38	39	40	41	42	43	44
Количество пар, n_i	1	2	3	5	10	13	9	6	1

64. В таблице приведены данные о выручке за сутки 200 торговых точек города. При $\alpha = 0{,}05$ проверить гипотезу о согласии выборочного распределения с нормальным законом.

Выручка за сутки, x_i (млн.руб.)	3,2	3,4	3,6	3,8	4	4,2	4,4	4,6	4,8	5
Количество торговых точек, n_i	2	5	4	18	83	62	14	6	4	2

65. В таблице дано распределение 1000 экземпляров северной сосны по диаметру ствола. При $\alpha = 0{,}05$ проверить гипотезу о том, что случайная величина X – диаметр ствола распределена по нормальному закону.

Диаметр ствола, см.	14-18	18-22	22-26	26-30	30-34
Количество сосен	16	35	109	183	214

Диаметр ств.	34-38	38-42	42-46	46-50	50-54	54-58
Кол-во сосен	197	115	71	36	19	5

66. Дано следующее распределение успеваемости 100 студентов-заочников, сдававших четыре экзамена

Число сданных экзаменов, x_i	0	1	2	3	4
Число студентов, n_i	1	1	3	35	60

Можно ли на уровне $\alpha = 0{,}01$ считать, что число сданных экзаменов среди четырех имеет биномиальный закон распределения?

67. В таблице приведены данные о количестве междугородних вызовов за 1 час 40 мин. Приняв уровень значимости $\alpha = 0{,}05$, проверить гипотезу о том, что распределение количества вызовов подчиняется закону Пуассона.

Количество вызовов за 1 мин., x_i	0	1	2	3	4	5
Частота, n_i	8	28	31	18	10	5

68. В продовольственном магазине сделаны контрольные замеры проданной колбасы. Отклонения от истинного веса даны в таблице

x_i, г.	1	2	3	4	5	6	7	8	9
n_i	16	5	17	18	18	11	6	29	20

При уровне значимости $\alpha = 0{,}05$ установить, согласуются ли данные выборки с законом равномерного распределения, приняв вероятность появления недовеса $p_i = 0{,}1,\ i = \overline{1{,}9}$.

Контрольные и тестовые задания

1. Статистическая гипотеза – это предложение о том, что:

а) средний балл по экономической теории в предстоящей аттестации будет выше 50;

б) рост мужчин определенной возрастной группы есть нормально распределенная случайная величина;

в) модным в летнем сезоне будет оливковый цвет джинсов;

г) две производственные линии дают равные доли (проценты) бракованной продукции;

д) предстоит повышение стипендии студентам ВУЗов.

2. Простая двухальтернативная параметрическая гипотеза – это:

а) $H_0 : MX = 2;\quad H_1 : MX = 0$;

б) $H_0 : X \sim N(\alpha, \sigma),\quad \alpha$ и σ - неизвестны;

в) $H_0 : DX = 1;\quad H_1 : DX = 2$; г) $H_0 : P(A) = 0{,}2;\quad H_1 : P(A) > 0{,}2$;

3. Цена некой ценной бумаги нормально распределена. По результатам наблюдений торгов за последние три месяца это предположение было отвергнуто:

а) принято верное решение;

б) допущена ошибка 1-го рода;

в) допущена ошибка 2-го рода.

4. Уровень значимости α при проверке статистических гипотез – это вероятность:

а) принять H_0 при условии, что H_0 верна;

б) отвергнуть H_0 при условии, что H_0 верна;

в) принять H_0 при условии, что H_0 неверна;

г) отвергнуть H_0 при условии, что H_0 неверна.

5. Стандартные значения уровня значимости α:

а) 0,05; б) 0,01; в) 0,90; г) 0,95.

6. Критическая область – это область значений статистики критерия, при попадании в которую:

а) H_0 принимается; б) H_0 отвергается;

в) решение принимается случайным образом.

7. Для параметрической гипотезы $H_0 : MX = 5;\quad H_1 : MX \neq 5$ критическая область:

а) правосторонняя; б) двусторонняя;

в) левосторонняя; г) любая из перечисленных выше.

8. Значение статистики критерия для проверки гипотезы $H_0 : DX = 0,2;\quad H_1 : DX > 0,2\quad \chi^2_{набл.} = 16,3;$ критическая точка $\chi^2_{кр.} = 18,5$. Принимается решение:

а) нет оснований отвергнуть H_0;

б) наблюдения не согласуются с H_0;

в) нет однозначного ответа.

9. Критерий согласия – это критерий проверки гипотез:

а) о равенстве средних;

б) о равенстве дисперсий;

в) о наличии корреляции между исследуемыми величинами;

г) о виде закона распределения.

10. При проверке гипотезы о нормальном распределении исследуемой случайной величины по критерию Пирсона (критерий χ^2) число интервалов группированной выборки равно 10. Число степеней свободы для статистики критерия равно:

а) 9; б) 8; в) 7; г) 6.

Приложение 1

Значения функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	0,3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	0,3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	0,3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	0,3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0,5	0,3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	0,3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	0,3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	0,2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	0,2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	0,2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	0,1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	0,1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	0,1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	0,1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	0,1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0,0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0,0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1,9	0,0656	0644	0632	0620	0608	0596	0584	0573	0562	0551

2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0,0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0,0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0,0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0,0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0,0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0,0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0,0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0,0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0,0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0,0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0,0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0,0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0,0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0,0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0,0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0,0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0,0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0,0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Приложение 2

Значения функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$

x	Сотые доли								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0200	0,0239	0,0279	0,0319
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810
1,2	0,3849	0,3869	0,3883	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761
2,0	0,4772	0,4779	0,4783	0,4788	0,4793	0,4798	0,4803	0,4807	0,4812
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4879	0,4881	0,4884	0,4887
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980
2,9	0,4981	0,4982	0,4983	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990

Приложение 3

Квантили распределения “Хи-квадрат” $\chi_p^2(k)$

p / k	**0,010**	**0,025**	**0,05**	**0,10**	**0,90**	**0,95**	**0,975**	**0,990**
1	0,0157	0,0982	0,0393	0,0158	2,71	3,84	5,02	6,63
2	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21
3	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3
4	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3
5	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1
6	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8
7	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5
8	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1
9	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7
10	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2
11	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7
12	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2
13	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7
14	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1
15	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6
16	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0
17	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4
18	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8
19	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2
20	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6
21	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9
22	9,54	11,0	12,3	14,0	30,8	33,9	36,8	40,3
23	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6
24	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0
25	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3
26	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6
27	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0
28	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3
29	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6
30	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9
35	18,5	20,6	22,5	24,8	46,1	49,8	53,2	57,3
40	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7
45	25,9	28,4	30,6	33,4	57,5	61,7	65,4	70,0
50	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2
75	49,5	52,9	56,1	59,8	91,1	96,2	100,8	106,4
100	70,1	74,2	77,9	82,4	118,5	124,3	129,6	135,6

Приложение 4

Квантили распределения Стьюдента $t_p(k)$

k \ p	**0,900**	**0,950**	**0,975**
1	3,078	6,314	12,706
2	1,886	2,920	4,303
3	1,638	2,353	3,182
4	1,533	2,132	2,776
5	1,476	2,015	2,571
6	1,440	1,943	2,447
7	1,415	1,895	2,365
8	1,397	1,860	2,306
9	1,383	1,833	2,262
10	1,372	1,812	2,228
11	1,363	1,796	2,201
12	1,356	1,782	2,179
13	1,350	1,771	2,160
14	1,345	1,761	2,145
15	1,341	1,753	2,131
16	1,337	1,746	2,120
17	1,333	1,740	2,110
18	1,330	1,734	2,101
19	1,328	1,729	2,093
20	1,325	1,725	2,086
21	1,323	1,721	2,080
22	1,321	1,717	2,074
23	1,319	1,714	2,069
24	1,318	1,711	2,064
25	1,316	1,708	2,060
26	1,315	1,706	2,056
27	1,314	1,703	2,052
28	1,313	1,701	2,048
29	1,311	1,699	2,045
30	1,310	1,697	2,042
40	1,303	1,684	2,021
60	1,296	1,671	2,000
120	1,289	1,658	1,980
∞	1,282	1,645	1,960

k_2 \ k_1	p = 0,95							
	9	**10**	**12**	**15**	**20**	**24**	**30**	**40**
1	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1
2	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47
3	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59
4	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72
5	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46
6	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77
7	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34
8	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04
9	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83
10	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66
11	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53
12	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43
13	2,71	2,67	2,60	2,63	2,46	2,42	2,38	2,34
14	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27
15	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20
16	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15
17	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10
18	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06
19	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03
20	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99
21	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96
22	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94
23	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91
24	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89
25	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87
26	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85
27	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84
28	2,24	2,19	2,12	2,04	1,96	1,941	1,87	1,82
29	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81
30	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79
40	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69
60	1,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59
120	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50
∞	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39

Рекомендуемая литература

1. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1975, 1997, 1998, 1999, 2000.
2. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1978, 1997, 1998, 1999, 2000.
3. Ежова Л.Н. и др. Теория вероятностей. Учебное пособие. Иркутск: Изд-во ИГЭА, 1996.
4. Ежова Л.Н. Теория вероятностей и математическая статистика. Иркутск: Изд-во ИГЭА, 2000.
5. Колемаев В.А., Староверов О.В., Турундаевский В.Б. Теория вероятностей и математическая статистика. – М.: Высш. Школа, 1991.
6. Колде Я.К. Практикум по теории вероятностей и математической статистики. М.: Высш. Школа, 1991.
7. Мацкевич И.П., Свирид Г.П., Булдык Г.М. Сборник задач и упражнений по высшей математике: Теория вероятностей и математическая статистика: Учеб. Пособие. Минск: Высш. Школа, 1996.