Вероятностные методы анализа и моделирования систем

Элементарные понятия о случайных событиях, величинах и функциях

Под *событием* понимается всякий факт, который может произойти в данных условиях. Теория вероятностей рассматривает события в тесной связи с теми условиями, в которых они наступают. Совокупность условий, в которых рассматривается данное событие, называют комплексом условий, а реализацию этого комплекса условий на практике – испытанием. В зависимости от связи между событиями и соответствующими комплексами условий различают *достоверные*, *невозможные* и *случайные* события. *Достоверным* называется такое событие, которое наступает каждый раз при реализации данного комплекса условий. Достоверное событие обозначим через U. *Невозможным* называется событие, которое никогда не наступает при реализации данного комплекса условий. Невозможное событие обозначим символом $\varnothing$. *Случайным* называется событие, которое может либо наступить при реализации данного комплекса условий, либо не наступить. Достоверное и невозможное события могут рассматриваться как крайние частные случаи случайных событий. Случайные события обозначим через A, B, C.

Элементарное событие – это один из нескольких возможных, но несовместных исходов того или иного опыта (испытания). Совокупность или множество их составляют пространство элементарных событий.

Всякому событию при данном комплексе условий соответствует определенная степень возможности. Более возможные события при многократных испытаниях в среднем наступают чаще, а менее возможные – реже. *Частотой события* называется отношение числа испытаний, в которых появилось данное событие, и общего числа испытаний. Частота события A равна:

$$P^*(A) = \frac{m(A)}{n},$$

где n – общее число проведенных испытаний; $m(A)$ – число испытаний, в которых наступило событие A.

Количественной мерой степени возможности появления события для заданного комплекса условий является *вероятность события*. Чем более возможно появление случайного события, тем больше его вероятность. Наоборот, чем менее возможно появление события, тем меньше его вероятность.

Случайные события могут быть представлены через случайные величины. Понятие «случайная величина» расширяет область применения вероятностных методов в решении практических задач, позволяет исследовать более сложные случайные явления. *Случайной* называется такая величина, которая в результате испытания (реализации определенного комплекса условий) может принять то или иное значение, причем до испытания неизвестно, какое именно. Если повторять испытания, то результатом каждого будет какое-либо одно значение случайной величины из множества возможных.

Случайные величины подразделяются на *дискретные* и *непрерывные*.

Множество значений дискретной случайной величины конечно или счетно, например:

- количество отказов автомобилей автопредприятия в течение рабочей смены;
- число рабочих, пришедших в бухгалтерию завода в течение одного часа получать заработную плату и т. д.

Множество значений непрерывной случайной величины представляет собой множество всех точек, принадлежащих какому-либо интервалу числовой оси, например:

- расход топлива на километр пробега;
- время безотказной работы автомобиля и т. д.

Для того чтобы задать случайную величину, необходимо задать множество значений, которые она может принимать. Однако одного перечня значений случайной величины еще недостаточно для каких-либо существенных выводов. Нужно еще знать, как часто, т.е. с какой вероятностью, она принимает эти значения. Ответ на поставленный вопрос дает исчерпывающая характеристика случайной величины – закон ее распределения.

Закон распределения представляет собой соотношение, позволяющее определить вероятность появления случайной величины в любом интервале (и, в частности, вероятности любых значений случайной величины).

Основными формами закона распределения являются: ряд распределения, функция распределения и плотность распределения.

Ряд распределения представляет собой таблицу, в которой перечислены возможные значения случайной величины и соответствующие им вероятности (табл. 2.1):

Таблица 2.1

Возможные значения случайной величины	x_1	x_2	...	x_n
Вероятности ($p_1 + p_2 + ... + p_n = 1$)	p_1	p_2	...	p_n

Эмпирический ряд распределения представляет собой таблицу, в которой перечислены наблюдаемые значения (фактические реализации) случайной величины и соответствующие им частоты (табл. 2.2).

Таблица 2.2

Возможные значения случайной величины	x_1	x_2	...	x_n
Частоты	m_1	m_2	...	m_n

Ряд распределения не может служить характеристикой непрерывной случайной величины, поскольку значения этой величины нельзя перечислить, так как множество их несчетно. Кроме того, вероятность отдельного значения непрерывной случайной величины равна нулю.

Поэтому ставят в соответствие вероятности не отдельные значения случайной величины, а множество значений ($X < x$), где x – произвольное число. Этот способ годится для дискретной и непрерывной случайных величин.

Функцией распределения (ФР) (или интегральным законом распределения) случайной величины X называется числовая функция

$F(x) = P\{X < x\}$, определенная для любых $x \in R$.

Свойства ФР:

1. $0 \le F(x) \le 1$;
2. $F(x_1) \le F(x_2)$, если $x_1 < x_2$, т.е. $F(x)$ – неубывающая функция;
3. $P\{a \le X < b\} = F(b) - F(a)$.

Плотностью распределения (ПР) (или дифференциальным законом распределения) случайной величины X называется числовая функция $f(x)$, равная производной от ФР, если такая производная существует: $f(x) = F'(x)$. Связь между ПР и ФР можно представить в интегральной форме:

$$\int_a^b f(x) \cdot dx = F(b) - F(a),$$

что позволяет определить ФР:

$$F(x) = \int_{-\infty}^{x} f(\tau) \cdot d\tau.$$

Свойства ПР:

1. $f(x) \ge 0$, т.к. ФР – неубывающая функция;
2. $P\{a \le X < b\} = \int_a^b f(x) \cdot dx$,
3. $\int_{-\infty}^{+\infty} f(\tau) \cdot d\tau = 1$ – условие нормировки.

Числовые характеристики случайных величин

Различают следующие группы числовых характеристик: *характеристики положения* (математическое ожидание, мода, медиана, квантиль и др.), *рассеивания* (дисперсия, среднеквадратичное отклонение и др.), *характеристики формы плотности распределения* (показатель асимметрии, эксцесса и др.).

Математическим ожиданием (средним значением по распределению) называется действительное число, определяемое в зависимости от типа случайной величины формулой:

$m_x = M[X] = a = \sum_{i=1}^{n} x_i p_i$ для дискретной случайной величины;

$m_x = M[X] = a = \int_{-\infty}^{+\infty} x f(x) dx$ для непрерывной случайной величины.

Математическое ожидание является теоретической характеристикой случайной величины. Эмпирической характеристикой случайной величины является *эмпирическая средняя*, вычисляемая по формуле

$$m^* = \frac{1}{N} \sum_{i=1}^{N} x_i \,.$$

Медианой Me случайной величины называется такая величина Me(X)= x_0, относительно которой равновероятно получение большего или меньшего значения случайной величины: $P\{X < x_0\} = P\{X \geq x_0\}$ или $F(x_0)$=0,5.

Модой Mo дискретной случайной величины называется ее значение, обладающее наибольшей вероятностью. Для непрерывной случайной величины мода есть такое значение, которое отвечает максимальной плотности распределения.

Дисперсией называется математическое ожидание квадрата отклонений случайной величины от своего математического ожидания:

$$D_x = \sigma_x^2 = M[(X–m_x)^2] = M[X^2] – m_x{}^2.$$

Дисперсия дискретной случайной величины равна

$$D_x = \sigma_x^2 = \sum_{i=1}^{n} \tilde{o}_i^2 p_i - m_x^2 .$$

Дисперсия непрерывной случайной величины равна

$$D_x = \sigma_x^2 = \int_{-\infty}^{+\infty} x^2 f(x) dx - m_x^2 .$$

При больших объемах выборки для эмпирической дисперсии используют формулу:

$$D^* = \frac{\sum_{i=1}^{N} (X_i - m^*)^2}{N} .$$

При относительно малых выборках ($N \leq 30$) следует пользоваться формулой для исправленной дисперсии:

$$S^2 = D^{**} = \frac{D^* N}{N-1} .$$

Наряду с дисперсией случайной величины, в качестве характеристики рассеивания случайной величины используется *среднее квадратическое отклонение*, которое равно положительному значению корня квадратного из дисперсии.

Относительной характеристикой рассеивания является *коэффициент вариации*, вычисляемый как отношение среднего квадратического отклонения и эмпирической средней

$$V = \frac{\sigma_x}{\overline{x}} \cdot 100\% .$$

Статистическая оценка законов распределения случайных величин

Полученная в результате статистического наблюдения *выборка* из n значений (*вариант*) изучаемого количественного *признака* X образует *вариационный ряд*. *Ранжированный* вариационный ряд получают, расположив варианты x_j, где $j = 1, 2, \ldots, n$, в порядке возрастания значений, то есть $x_1 \le x_2 \le \ldots \le x_j \le \ldots \le x_n$.

Изучаемый признак X может быть *дискретным*, то есть его значения отличаются на конечную, заранее известную величину (год рождения, тарифный разряд, число людей), или *непрерывным*, то есть его значения отличаются на сколь угодно малую величину (время, вес, объем, стоимость).

Частотой m_i в случае дискретного признака X называют число одинаковых вариант x_i, содержащихся в выборке. В ранжированном вариационном ряду одинаковые варианты очевидно расположены подряд:

$$\overbrace{\underbrace{x_1, x_1, \ldots, x_1}_{m_1}, \ldots, \underbrace{x_i, x_i, \ldots, x_i}_{m_i}, \ldots, \underbrace{x_k, x_k, \ldots, x_k}_{m_k}}^{n} .$$

Вариационный ряд для дискретного признака X принято наглядно и компактно представлять в виде таблицы, в первой строке которой указаны k различных значений x_i изучаемого признака, а во второй строке – соответствующие этим значениям частоты m_i, где $i = 1, 2, \ldots, k$. Такую таблицу называют *статистическим (выборочным) распределением*.

Статистическое распределение для непрерывного признака X принято представлять *интервальным рядом* – таблицей, в первой строке которой указаны k интервалов значений изучаемого признака X вида $(x_{i-1} – x_i)$, а во второй строке – соответствующие этим интервалам частоты m_i, где $i = 1, 2, \ldots, k$. Обозначение $(x_{i-1} - x_i)$ – указывает не разности, а все значения признака X от x_{i-1} до x_i, кроме правой границы интервала x_i.

Для непрерывного признака X *частота* m_i – число различных x_j, попавших в соответствующий интервал: $x_j \in [x_{i-1}\, ; x_i)$.

Вместо частот m_i во второй строке могут быть указаны *относительные частоты* $w_i = \dfrac{m_i}{n}$ (*частости*). Очевидно, что сумма частот равна объему выборки (выборочной совокупности) n , а сумма относительных частот (частостей) равна единице:

$$\sum_{i=1}^{k} m_i = n, \qquad \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{m_i}{n} = 1.$$

Далее показаны четыре возможных формы представления статистических распределений с соответствующими краткими названиями:

Дискретный ряд частот

x_i	x_1	x_2	…	x_k
m_i	m_1	m_2		m_k ,

Интервальный ряд частот

x_{i-1}–x_i	x_0–x_1	x_1–x_2	…	x_{k-1}–x_k
m_i	m_1	m_2	…	m_k ,

Дискретный ряд частостей

x_i	x_1	x_2	…	x_k
w_i	w_1	w_2		w_k ,

Интервальный ряд частостей

x_{i-1}–x_i	x_0–x_1	x_1–x_2	…	x_{k-1}–x_k
w_i	w_1	w_2	…	w_k .

По аналогии с *теоретической* функцией распределения генеральной совокупности $F(x)$, которая определяет вероятность события $X < x$: $F(x) = P(X < x)$, вводят понятие *эмпирической функции распределения* $F^*(x)$, которая определяет *относительную частоту* этого же события $X < x$, то есть $F^*(x) = \dfrac{m(X < x)}{n}$.

Для наглядности принято использовать следующие формы графического представления статистических распределений:

- дискретный ряд изображают в виде полигона. *Полигон частот* – ломаная линия, отрезки которой соединяют точки с координатами (x_i , m_i); аналогично, *полигон относительных частот* – ломаная, отрезки которой соединяют точки с координатами (x_i, w_i);
- интервальный ряд изображают в виде гистограммы. *Гистограмма частот* есть ступенчатая фигура, состоящая из прямоугольников, основания которых – интервалы длиной h_i, а высоты – плотности частот $\dfrac{m_i}{h_i}$. В случае *гистограммы относительных частот* высоты прямоугольников – плотности относительных частот $\dfrac{w_i}{h_i} = \dfrac{m_i}{n \cdot h_i}$. Здесь в общем случае $h_i = x_i - x_{i-1}$, однако на практике чаще всего полагают величину h одинаковой для всех интервалов:

$h_i = h = (x_k - x_0)/k$, где $i = 1,2,\dots k$. Очевидно для ранжированного вариационного ряда $x_0 \le x_{\min} = x_{(1)}$; $x_k > x_{\max} = x_{(n)}$. В скобках указаны индексы j исходного ранжированного вариационного ряда.

Замечание 1. Если в статистическом исследовании исходным является статистическое распределение в виде интервального ряда (сгруппированные данные), а исходный вариационный ряд недоступен, то точное расположение отдельных вариант, попавших в каждый из интервалов неизвестно. Только выбирая в качестве аргумента эмпирической функции распределения правую границу интервала (x_{i-1}–x_i), мы уверены, что все варианты, попавшие в этот интервал, будут учтены (просуммированы) в значении накопленной частоты (накопленной относительной частоты), соответствующей этому интервалу.

Пример. Вариационный ряд часовой выработки автомобиля КамАЗ-5511 имеет вид:

Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
Частота	7	14	17	17	15	14	11	5

Построить гистограмму и статистическую функцию распределения часовой выработки подвижного состава автопредприятия.

Решение.

Высоты прямоугольников гистограммы – плотности относительных частот $\dfrac{w_i}{h} = \dfrac{m_i}{n \cdot h}$.

$$h = x_2 - x_1 = 5,5 - 4 = 1,5.$$

Отсюда находим:

1) $\dfrac{m_1}{n \cdot h} = \dfrac{7}{100 \cdot 1,5} = 0,047$;

2) $\dfrac{m_2}{n \cdot h} = \dfrac{14}{100 \cdot 1,5} = 0,093$;

3) $\dfrac{m_3}{n \cdot h} = \dfrac{17}{100 \cdot 1,5} = 0,113$;

4) $\dfrac{m_4}{n \cdot h} = \dfrac{17}{100 \cdot 1,5} = 0,113$;

5) $\dfrac{m_5}{n \cdot h} = \dfrac{15}{100 \cdot 1,5} = 0,1$

6) $\dfrac{m_6}{n \cdot h} = \dfrac{14}{100 \cdot 1,5} = 0,093$

7) $\dfrac{m_7}{n \cdot h} = \dfrac{11}{100 \cdot 1,5} = 0,073$

8) $\dfrac{m_8}{n \cdot h} = \dfrac{17}{100 \cdot 1,5} = 0,033$

Построим диаграмму (рис. 2.1)

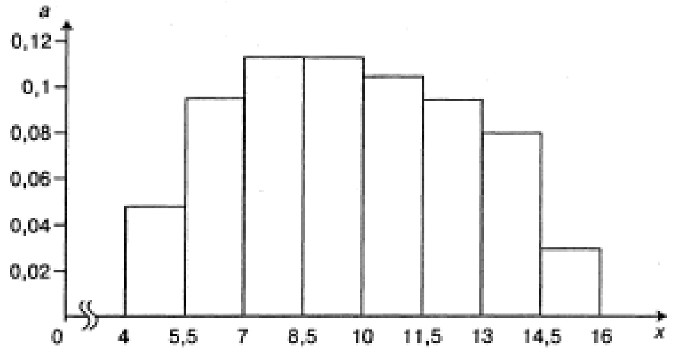


Рис. 2.1. Гистограмма часовой выработки автомобиля

Интервальный ряд относительных частот часовой выработки автомобиля КамАЗ-5511 имеет вид:

Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
Частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05

Построим статистическую функцию распределения часовой выработки автомобиля:

1) при $x \leq 4$ $F^*(x)$=0;

2) при 4 < $x \leq 5{,}5$ $F^*(x)$=0,07;

3) при 5,5 < $x \leq 7$ $F^*(x)$=0,21;

4) при 7 < $x \leq 8{,}5$ $F^*(x)$=0,38;

5) при 8,5 < $x \leq 10$ $F^*(x)$=0,55;

6) при 10 < $x \leq 11{,}5$ $F^*(x)$=0,70;

7) при 11,5 < $x \leq 13$ $F^*(x)$=0,84;

8) при 13 < $x \leq 14{,}5$ $F^*(x)$=0,95;

9) при 14,5 < $x \leq 16$ $F^*(x)$=1;

График статистической функции распределения представлен на рис. 2.2.

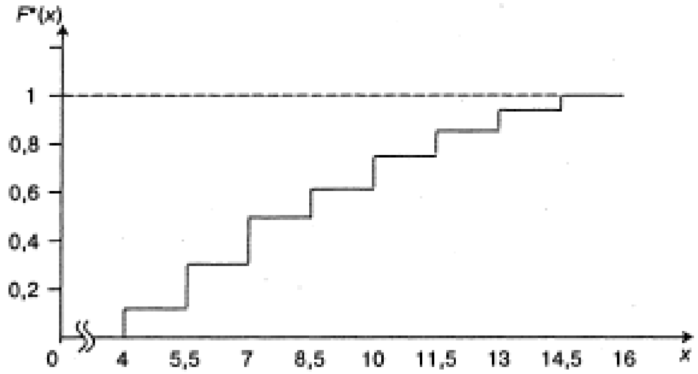


Рис. 2.2. Статистическая функция распределения часовой выработки автомобиля

Пример. Статистическое распределение выборки имеет вид

x_i	2	3	4	5
n_i	4	7	5	4

Тогда относительная частота варианты x=2 равна…?

Решение. n=4+7+5+4=20; тогда относительная частота варианты x=2 равна $w_1=n_1/n$=4/20=0,2.

Основные виды теоретических распределений

Равномерное распределение.

Непрерывная случайная величина имеет *равномерное распределение* на отрезке [a, b], если на этом отрезке плотность распределения случайной величины постоянна, а вне его равна нулю.

$$f(x)=\begin{cases}0, & \text{при } x<a;\\ \dfrac{1}{b-a}, & \text{при } a\le x\le b;\\ 0, & \text{при } x>b.\end{cases}$$

Кривая равномерного распределения показана на рис. 2.3.

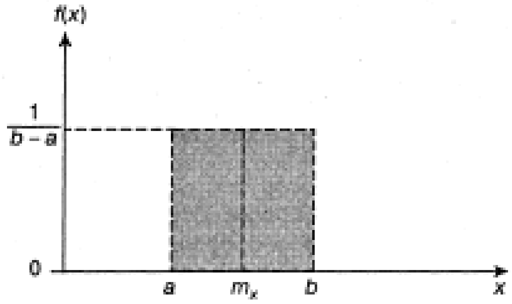


Рис. 2.3. Кривая равномерного распределения

Функция распределения $F(x)$ на отрезке [a,b] имеет вид:

$$F(x)=\begin{cases}0, & \text{при } x<a;\\ \dfrac{x-a}{b-a}, & \text{при } a\le x\le b;\\ 1, & \text{при } x>b.\end{cases}$$

График функции равномерного распределения представлен на рис. 2.4.

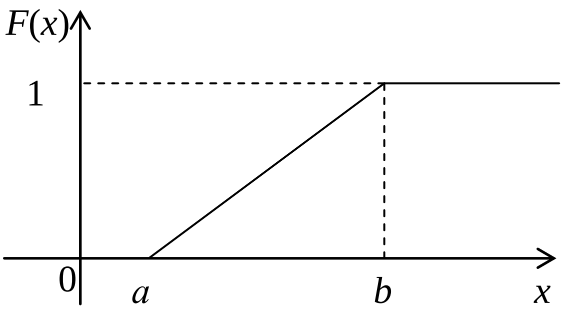


Рис. 2.4. График функции равномерного распределения

Математическое ожидание, дисперсия и среднее квадратическое отклонение случайной величины, подчиненной равномерному закону распределения равны, соответственно:

$$m_x = \frac{a+b}{2}; \qquad D_x = \frac{(b-a)^2}{12}; \qquad \sigma_x = \sqrt{D_x} = \frac{b-a}{2\sqrt{3}}.$$

Вероятность попадания случайной величины в заданный интервал:

$$P(\alpha < X < \beta) = \frac{\beta - \alpha}{b - a}.$$

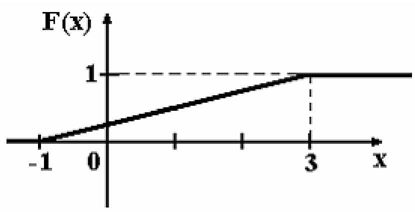
Пример. Троллейбусы прибывают на остановку через 4 мин. Какова вероятность того, что время ожидания троллейбуса не превысит 3 мин?

Решение.

Так как (β–α)=3 мин, а (b–a)=4 мин, то

$$P(0 < X < 3) = \frac{3}{4} = 0{,}75.$$

Пример. График функции распределения вероятностей непрерывной случайной величины X, распределенной равномерно в интервале (–1; 3), имеет вид:



Найти математическое ожидание.

Решение.

a= –1, b=3. Тогда $m_x = \dfrac{a+b}{2} = \dfrac{3-1}{2} = 1.$

Показательное распределение.

Показательным (экспоненциальным) называется распределение вероятностей непрерывной случайной величины X, которое описывается плотностью

$$f(x) = \begin{cases} 0, & \text{при} \quad x < 0; \\ \lambda e^{-\lambda x}, & \text{при} \quad x \ge 0. \end{cases}$$

где λ – положительное число.

Функция распределения имеет вид:

$$F(x) = \begin{cases} 0, & \text{при} \quad x < 0; \\ 1 - e^{-\lambda x} & \text{при} \quad x \ge 0. \end{cases}$$

Графики функции распределения и плотности распределения приведены на рис. 2.5 и рис. 2.6, соответственно:

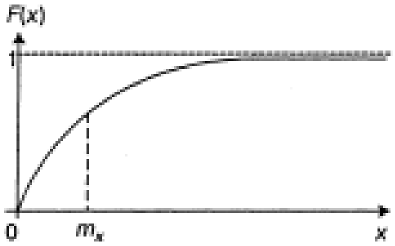


Рис. 2.5. График функции показательного распределения

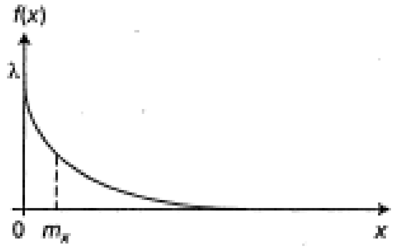


Рис. 2.6. График плотности показательного распределения

Математическое ожидание, дисперсия и среднее квадратическое отклонение случайной величины, подчиненной показательному закону распределения равны, соответственно:

$$M(X)=\frac{1}{\lambda};\quad D(X)=\frac{1}{\lambda^2};\quad \sigma_x=\frac{1}{\lambda}.$$

Пример. Случайная величина X распределена по показательному закону, ее плотность распределения вероятностей имеет вид

$$f(x)=\begin{cases}0, & x<0,\\ 1,3e^{-1,3x}, & x\ge 0.\end{cases}$$

Найти математическое ожидание, стандартное отклонение, дисперсию.

Решение.

$$M(X)=\frac{1}{\lambda}=\frac{1}{1,3}=0,77;\quad D(X)=\frac{1}{\lambda^2}=\frac{1}{1,3^2}=0,59;\quad \sigma_x=\frac{1}{\lambda}=\frac{1}{1,3}=0,77.$$

Также легко определить и вероятность попадания случайной величины, подчиненной показательному закону распределения, в заданный интервал.

$$P(a<x<b)=F(b)-F(a)=e^{-\lambda a}-e^{-\lambda b}.$$

Показательное распределение широко используется в теории надежности.

Допустим, некоторое устройство начинает работать в момент времени $t_0=0$, а через какое– то время t происходит отказ устройства.

Обозначим T непрерывную случайную величину – длительность безотказной работы устройства.

Таким образом, функция распределения $F(t) = P(T<t)$ определяет вероятность отказа за время длительностью t.

Вероятность противоположного события (безотказная работа в течение времени t) равна $R(t) = P(T>t) = 1 – F(t)$.

Функцией надежности $R(t)$ называют функцию, определяющую вероятность безотказной работы устройства в течение времени t.

Часто на практике длительность безотказной работы подчиняется показательному закону распределению.

Вообще говоря, если рассматривать новое устройство, то вероятность отказа в начале его функционирования будет больше, затем количество отказов снизится и будет некоторое время иметь практически одно и то же значение. Затем (когда устройство выработает свой ресурс) количество отказов будет возрастать.

Другими словами, можно сказать, что функционирование устройства на протяжении всего существования (в смысле количества отказов) можно описать комбинацией двух показательных законов (в начале и конце функционирования) и равномерного закона распределения.

Функция надежности для какого-либо устройства при показательном законе распределения равна:

$$R(t) = 1 - F(t) = e^{-\lambda t}.$$

Данное соотношение называют *показательным законом надежности*.

Важным свойством, позволяющим значительно упростить решение задач теории надежности, является то, что вероятность безотказной работы устройства на интервале времени t не зависит от времени предшествующей работы до начала рассматриваемого интервала, а зависит только от длительности времени t.

Таким образом, безотказная работа устройства зависит только от интенсивности отказов λ и не зависит от безотказной работы устройства в прошлом.

Нормальный закон распределения.

Нормальным называется распределение вероятностей непрерывной случайной величины, которое описывается плотностью вероятности

$$f(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}}.$$

Нормальный закон распределения также называется *законом Гаусса*.

Параметры m_x и σ_x, входящие в плотность распределения являются соответственно математическим ожиданием и средним квадратическим отклонением случайной величины X.

Пример. Случайная величина X подчинена нормальному закону распределения. Плотность вероятности $f(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-3)^2}{8}}$. Найти математическое ожидание случайной величины X.

Решение. $m_x = 3$.

Найдем функцию распределения $F(x)$.

$$F(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-m_x)^2}{2\sigma_x^2}} dx.$$

График плотности нормального распределения (рис. 2.7) называется *нормальной кривой* или *кривой Гаусса*.

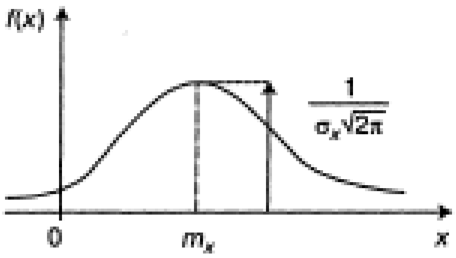


Рис. 2.7. График плотности нормального распределения

Распределение Вейбулла.

Распределение Вейбулла описывает явления, связанные с задачами долговечности и усталости, в теории хрупкого разрушения материалов и др. Плотность распределения имеет вид:

$$f(x) = \frac{a}{b}\left(\frac{x}{a}\right)^{a-1} e^{-\left(\frac{x}{b}\right)^{a}}.$$

Выбор теоретического закона распределения случайной величины

Подбор подходящего теоретического распределения. Критерии согласия

При наличии числовых характеристик случайной величины (математического ожидания, дисперсии, коэффициента вариации) законы ее распределения могут быть определены в первом приближении по таблице 2.3.

Таблица 2.3

Законы распределения случайной величины в зависимости от коэффициента вариации

Пределы изменения коэффициента вариации V_x	Закон распределения случайной величины X
$V_x \le 0{,}3$	Нормальный
$0{,}3 \le V_x < 0{,}4$	Гамма-распределение
$0{,}4 \le V_x < 1$	Вейбулла
$V_x = 1$	Экспоненциальный, Пуассона

Чтобы подобрать подходящее теоретическое распределение, необходимо построить кривую плотности распределения, после этого выбрать похожую из известных типов распределений. Если есть основания отдать предпочтение тому или иному распределению, то кривую строить нет необходимости. Затем выдвигают гипотезу о соответствии экспериментального и теоретического распределений, проверяют её на заданном уровне значимости, используя критерии согласия. Существуют несколько критериев.

Критерий Пирсона (хи-квадрат) применим только к сгруппированным данным. Рекомендуется, чтобы объем выборки был больше 100 и численность интервалов (групп), была не менее 5. Исходные данные разбивают на m интервалов и вычисляют для каждого:

- *экспериментальные частоты* $p_i^* = n_i/n$, n_i – количество данных попавших в i – й интервал, n – объём выборки;
- *теоретические частоты* $p_i = F(x_{i+1}) - F(x_i)$, найденные по таблицам и формулам для выбранного типа теоретического распределения;
- *экспериментальную величину*

$$(\chi^2)^* = n\sum_{i=1}^{m}\frac{(p_i^* - p_i)^2}{p_i} = \sum_{i=1}^{m}\frac{(n_i - np_i)^2}{np_i}.$$

По таблицам квантилей распределения χ^2 при заданном уровне значимости β (обычно 5%) и известном числе степеней свободы f находят теоретическое значение χ^2. f равно количеству интервалов минус число независимых условий, наложенных на экспериментальные частоты p_i^*. Примерами таких условий могут быть: равенство единице суммы всех частот, совпадение статистического среднего с гипотетическим, совпадение дисперсий и т.п. Следовательно:

$$f=m-r-1,$$

где m – число интервалов, r – число параметров, определяемых из опытных данных.

Пример. Если предполагаемое распределение – нормальное, то оценивают два параметра – математическое ожидание и среднее квадратическое отклонение, тогда $f=m-r-1=m-2-1=m-3$.

Если $(\chi^2)^*<\chi^2$, то функция распределения при заданном уровне значимости (β=5%) согласуется с экспериментальными данными.

Пример. Пользуясь критерием Пирсона, подобрать теоретический закон распределения для часовой выработки автомобилей КамАЗ-5511, статистическое распределение которой приведено в таблице 2.4.

Таблица 2.4

Вариационный ряд часовой выработки автомобиля

Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
Отн.частота	0,07	0,14	0,17	0,17	0,15	0,14	0,11	0,05

По форме гистограммы рис.2.8 можно предположить, что часовая выработка автомобиля подчиняется нормальному закону.

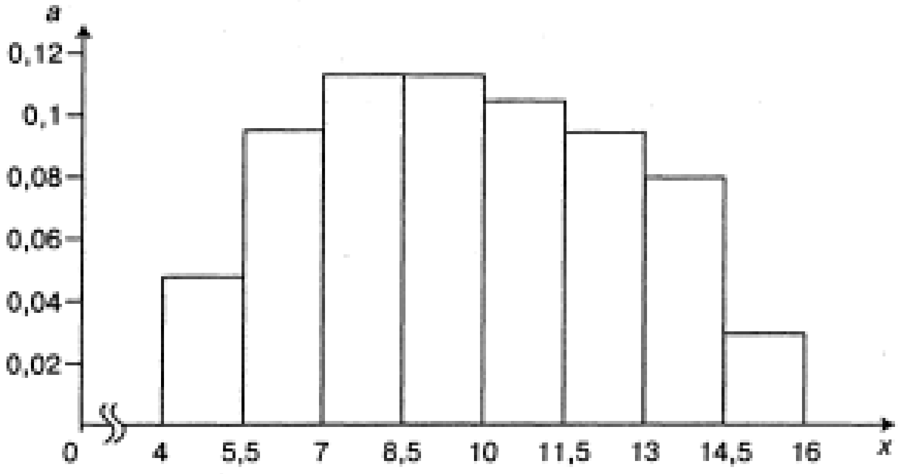


Рис. 2.8. Гистограмма часовой выработки автомобиля

Для оценки числовых характеристик нормального распределения вычислим:

математическое ожидание

$$m_x = \sum_{i=1}^{k} \overline{x}_i p_i^* = 4,75 \cdot 0,07 + 6,25 \cdot 0,14 + 7,75 \cdot 0,17 + 9,25 \cdot 0,17 +$$

$$+10,75 \cdot 0,15 + 12,25 \cdot 0,14 + 13,75 \cdot 0,11 + 15,25 \cdot 0,05 = 9,7 \text{ ò};$$

дисперсию

$$D_x = \sum_{i=1}^{k} (m_x - \overline{x}_i) p_i^* = (9,7 - 4,75) \cdot 0,07 + (9,7 - 6,25) \cdot 0,14 +$$

$$+(9,7 - 7,75) \cdot 0,17 + (9,7 - 9,25) \cdot 0,17 + (9,7 - 10,75) \cdot 0,15 +$$

$$+(9,7 - 12,25) \cdot 0,14 + (9,7 - 13,75) \cdot 0,11 + (9,7 - 15,25) \cdot 0,05 = 8,48;$$

среднее квадратическое отклонение

$$\sigma_x = \sqrt{D_x} = \sqrt{8,48} = 2,91;$$

коэффициент вариации

$$V_x = \frac{\sigma_x}{m_x} = \frac{2,91}{9,7} = 0,3.$$

Величина V_x=0,3 свидетельствует о том, что теоретическое распределение близко к нормальному закону распределения. Проверим данную гипотезу, воспользовавшись критерием согласия Пирсона.

Определим теоретическую вероятность попадания значений часовой выработки автомобиля в заданные интервалы, используя формулу:

$$p_i = F(x_{i+1}) - F(x_i) = \varPhi\left(\frac{x_{i+1} - m_x}{\sigma_x}\right) - \varPhi\left(\frac{x_i - m_x}{\sigma_x}\right),$$

где x_i, x_{i+1} – границы i-го интервала (табл. 2.4),

$\varPhi(u)$ – функция Лапласа (приложение 1).

Составим расчетную таблицу 2.5.

Таблица 2.5

x_i	x_{i+1}	u_i	u_{i+1}	$\Phi(u_i)$	$\Phi(u_{i+1})$	p_i	np_i	np_i
4	5,5		-1,44	-0,5	-0,4251	0,075	7,5	7
5,5	7	-1,44	-0,93	-0,4251	-0,3238	0,101	10,1	10
7	8,5	-0,93	-0,41	-0,3238	-0,1591	0,165	16,5	17
8,5	10	-0,41	0,10	-0,1591	0,0398	0,199	19,9	20
10	11,5	0,10	0,62	0,0398	0,2324	0,193	19,3	19
11,5	13	0,62	1,13	0,2324	0,3708	0,138	13,8	14
13	14,5	1,13	1,65	0,3708	0,4505	0,080	8,0	8
14,5	16	1,65		0,4505	0,5	0,050	5,0	5

Затем составим сравнительную таблицу (табл. 2.6) чисел попаданий в интервалы m_i и соответствующих значений np_i (n=100).

Таблица 2.6

Сравнительная таблица

Интервал	4-5,5	5,5-7	7-8,5	8,5-10	10-11,5	11,5-13	13-14,5	14,5-16
Количество наблюдений, n_i	7	14	17	17	15	14	11	5
Теоретическое количество наблюдений, np_i	7	10	17	20	19	14	8	5

Построим график теоретического распределения и совместим его с гистограммой статистического распределения (рис. 2.9).

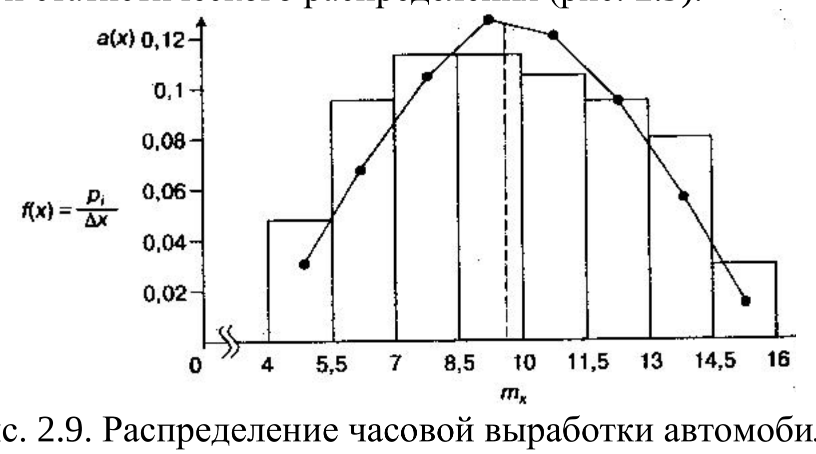


Рис. 2.9. Распределение часовой выработки автомобиля

Вычислим значение меры расхождения (табл. 2.7) по формуле

$$(\chi^2)^* = \sum_{i=1}^{8} \frac{(n_i - np_i)^2}{np_i} = 4,02 .$$

Таблица 2.7

n_i	np_i	$n_i - np_i$	$(n_i - np_i)^2$	$(n_i - np_i)^2 / np_i$
7	7	0,00	0	0
14	10	4,00	16	1,6
17	17	0,00	0	0
17	20	-3,00	9	0,45
15	19	-4,00	16	0,84
14	14	0,00	0	0
11	8	3,00	9	1,125
5	5	0,00	0	0
$\sum$=100	$\sum$=100			$\sum$=**4,02**

Определим число степеней свободы $f=m-r-1=8-2-1=5$.

По таблицам квантилей распределения χ^2 при заданном уровне значимости β=5% и числе степеней свободы f=5 находим теоретическое значение χ^2=11,1 (приложение 2).

Так как $(\chi^2)^* < \chi^2$, то функция распределения при заданном уровне значимости (β=5%) согласуется с экспериментальными данными и гипотезу о том, что часовая выработка автомобиля распределена по нормальному закону, можно считать правдоподобной.

Критерий Колмогорова-Смирнова определяется разностью максимальных абсолютных значений статистической функции распределения $F^*(x)$ и соответствующей теоретической функцией распределения $F(x)$, т.е.: $D = \max\left|F^*(x) - F(x)\right|$.

Было доказано, что какой бы вид не имела $F(x)$ при неограниченном возрастании числа независимых наблюдений n вероятность неравенства:

$$D\sqrt{n} \geq \lambda$$

стремится к пределу:

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}.$$

Для практического использования есть таблица квантилей, определённых из соотношения $k(\lambda_\alpha)=\alpha$, где функция распределения записана в несколько ином виде:

$$k(\lambda) = 1 - 2\sum_{\nu=1}^{\infty} (-1)^{\nu-1} e^{-2\nu^2\lambda^2}.$$

Схема применения критерия:

1) по результатам n наблюдений построить статистическую функцию распределения $F^*(x)$;
2) на том же графике нанести предполагаемую теоретическую функцию распределения $F(x)$;

3) определить max величину модуля разности ординат D и вычислить величину $\lambda = D\sqrt{n}$;
4) с помощью таблицы по заданному уровню значимости β найти значение λ_α. Если $\lambda < \lambda_\alpha$, то теоретическое и экспериментальное распределения согласуются на заданном уровне значимости.

Пример. Пусть задан массив случайных чисел $x_1, x_2, \ldots, x_n$. Это могут быть значения снеговой нагрузки, определенные по данным многолетних наблюдений, пределы текучести стали, найденные путем испытаний, усилия в элементе конструкции, вычисленные путем многолетних расчетов. Определить теоретическое распределение этих чисел.

Запишем исходный массив (табл. 2.8)

Таблица 2.8

№ п/п	x_i	№ п/п	x_i	№ п/п	x_i
1	110	16	112	31	116
2	99	17	112	32	116
3	103	18	112	33	116
4	127	19	112	34	108
5	106	20	111	35	117
6	116	21	113	36	119
7	109	22	113	37	110
8	92	23	113	38	111
9	110	24	114	39	121
10	125	25	114	40	122
11	113	26	114	41	125
12	125	27	114	42	111
13	111	28	115	43	120
14	121	29	115	44	126
15	129	30	116	45	105
				46	111

Прежде всего массив следует выстроить в порядке возрастания значений случайных величин (ранжировать) с указанием, сколько раз такое значение повторяется (табл. 2.9):

Таблица 2.9

x_i	92	99	103	105	106	108	109	110	111	112	113	114
n_i	1	1	1	1	1	1	1	3	5	4	4	4

Продолжение таблицы 2.9

x_i	115	116	117	119	120	121	122	125	126	127	129
n_i	2	5	1	1	1	2	1	3	1	1	1

Далее весь диапазон делят на интервалы, назначают их ширину и определяют количество (зависит от размера массива случайных величин, при больших массивах выбирают порядка 12…20 интервалов, а при физическом моделировании ограничиваются 6…12 интервалами). В данном примере берем 10 интервалов. Далее производим вычисления и заносим их в таблицу 2.10. В первом столбце – номера интервалов, во втором – границы, в третьем – среднее значение интервалов, в четвертом – количество случайных величин, попавших в интервал, причем, если случайные величины расположены на границе интервала, то половину из них относят к левому, а половину к правому интервалам, остальные столбцы понятны по их названиям:

Таблица 2.10

i	интервал	x_i	n_i	x_in_i	x_i^2	$x_i^2n_i$	$h_i=n_i/n$
1	92–96	94	1	94	8836	8836	0,022
2	96–100	98	1	98	9604	9604	0,022
3	100–104	102	1	102	10404	10404	0,022
4	104–108	106	2,5	265	11236	28090	0,054
5	108–112	110	11,5	1265	12100	139150	0,250
6	112–116	114	14,5	1653	12996	188442	0,315
7	116–120	118	5	590	13924	69620	0,109
8	120–124	122	3,5	427	14884	52094	0,076
9	124–128	126	5	630	15876	79380	0,109
10	128–132	130	1	130	16900	16900	0,022
	сумма		46	5254		602520	

По полученным данным строим гистограмму (рис. 2.10). На оси абсцисс откладываем интервалы (второй столбец), по оси ординат – частоты (8 столбец). Для построения экспериментальной плотности распределения должны иметь значение h_i/d, где d – величина интервала. В этом случае площадь, ограниченная гистограммой будет равна 1. Соединим вершины столбиков и получим экспериментальную кривую плотности распределения заданного массива. Далее, подбираем подходящее теоретическое распределение.

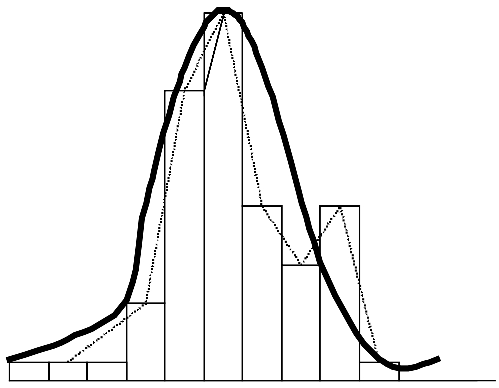


Рис. 2.10. Гистограмма плотности распределения

Определим статистические параметры распределения:

$$m_x = \frac{\sum x_i n_i}{n} = \frac{5254}{46} = 114,2;\; M(x^2) = \frac{\sum x_i^{\,2} n_i}{n} = 602520 / 46 = 13098;$$

$$D_x = M(x^2) - m_x^{\,2} = 13098 - 14046 = 52;$$

$$\sigma_x = \sqrt{D_x} = \sqrt{52} = 7,2;$$

$$V_x = \frac{\sigma_x}{m_x} = \frac{7,2}{114,2} = 0,06.$$

Величина V_x=0,06 свидетельствует о том, что теоретическое распределение близко к нормальному закону распределения. Проверим данную гипотезу, воспользовавшись критерием согласия Колмогорова-Смирнова.

Выразим функцию распределения через нормированную нормальную функцию распределения:

$$\Phi(u) = \Phi\left(\frac{x - m_x}{\sigma_x}\right),$$

что позволяет определить теоретические накопленные частоты p_i по данным таблицы содержащей интегральную функцию нормального распределения. При этом

$$p_i = 0,5 - \Phi(u) \text{ для } u < 0;$$

$$p_i = 0,5 + \Phi(u) \text{ для } u > 0.$$

Данные занесем в таблицу 2.11.

Анализируя шестой столбец, найдем максимальную разность теоретической и экспериментальной накопленных частот D=0,086, определяем $\lambda = D\sqrt{n} = 0,086\sqrt{46} = 0,583$.

Таблица 2.11

x_i	$(x_i – m_x)/\sigma$	$\Phi(u)$	p_i	H_i	$D=p_i–H_i$
1	2	3	4	5	6
92	–3,062	0,4987	0,001	0	0,001
96	–2,511	0,4938	0,006	0,022	-0,016
100	–1,959	0,475	0,025	0,044	-0,019
104	–1,408	0,4207	0,079	0,066	0,013
108	–0,857	0,3051	0,195	0,12	0,075
112	–0,306	0,1217	0,378	0,37	0,008
116	+0,246	0,0987	0,599	0,685	-0,086
120	+0,797	0,2881	0,788	0,794	-0,006
124	+1,348	0,4115	0,912	0,87	0,042
128	+1,899	0,4713	0,971	0,979	-0,008
132	+2,451	0,4931	0,993	1	-0,007

По таблице «Критические значения наибольшего отклонения эмпирического распределения от теоретического» (Приложение 3) найдем при уровне значимости 5% (α=0,95), $\lambda_\alpha = 1{,}34$. Так как $\lambda < \lambda_\alpha$, то по критерию Колмогорова-Смирнова разница между теоретическим и экспериментальным распределениями незначительная и нет оснований отвергать гипотезу о нормальном распределении совокупности.

Сравнение двух групп случайных величин

Часто бывает необходимо сравнить две выборки, чтобы отнести их к одной генеральной совокупности. Например, при оценке возможности рассматривать результаты наблюдений двух метеостанций в составе одного района. Такие задачи решаются на основе статистических проверок статистических гипотез.

Рассмотрим критерии Вилкоксона для проверки гипотезы об однородности двух выборок. Этот критерий применим к случайным величинам, распределения которых неизвестны, единственное требование, чтобы величины были непрерывны.

Пусть заданны выборки с объёмами n_1 и n_2, причём $n_1 \le n_2$, их функции распределения $F_1(x)$ и $F_2(x)$ – неизвестны. Требуется проверить на уровне значимости $\beta=2Q$ нулевую гипотезу, состоящую в том, что при всех значениях аргумента функции распределения равны между собой, то есть H_0: $F_1(x)=F_2(x)$ при конкурирующей гипотезе H_1: $F_1(x) \neq F_2(x)$.

Пусть объём выборки не превышает 25.

Этапы критерия:

1) расположить случайные величины обеих выборок в возрастающем порядке, то есть в виде одного вариационного ряда, и найти в этом ряду наблюдаемое значение критерия W – сумму порядковых номеров вариант первой выборки;
2) найти по таблице «критические точки критерия Вилкоксона» нижнюю критическую точку $w_n=f(Q,n_1,n_2)$, $Q=\beta/2$;
3) найти верхнюю критическую точку по формуле:
$$w_v = (n_1+n_2+1)n_1 - w_n;$$
4) Если $W < w_n$ или $W > w_v$ – H_0 отвергают, если $w_n < W < w_v$ нет оснований отвергать H_0.

Если объём выборок превышает 25, то для определения нижней критической точки следует использовать целую часть выражения:

$$w(Q,n_1,n_2) = \left[\frac{(n_1+n_2+1)n_1-1}{2} - Z_{ñr}\sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}\right],$$

где $Q=\beta/2$; Z_{cr} найти по таблице функции Лапласа по равенству

$$\Phi(Z_{cr})=(1-\beta)/2.$$

Пример. При уровне значимости 0,05 проверить H_0 об однородности двух выборок объёмом n_1=6, n_2=8 при H_1: $F_1(x) \neq F_2(x)$.

$x_{1,i}$	**15**	**23**	**25**	**26**	**28**	**29**		
$x_{2,i}$	12	14	18	20	22	24	27	30

Решение:

Расположим варианты обеих выборок в виде одного вариационного ряда и перенумеруем их:

№ п/п	1	2	**3**	4	5	6	**7**	8	**9**	**10**	11	**12**	**13**	14
варианты	12	14	**15**	18	20	22	**23**	24	**25**	**26**	27	**28**	**29**	30

Найдем наблюдаемое значение критерия Вилкоксона – сумму порядковых номеров по варианту первой выборки:

$$W=3+7+9+10+12+13=54.$$

Учитывая, что $Q=\beta/2=0{,}05/2=0{,}025$, n_1=6, n_2=8 из таблицы «критические точки критерия Вилкоксона» (приложение 3) найдем нижнюю критическую точку:

$$w_n\ (0{,}025,\ 6,\ 8)=29.$$

Найдем верхнюю критическую точку:

$$w_v=(n_1+n_2+1)n_1-w_n=(6+8+1)\cdot 6-29=61.$$

Так как 29<54<61, т.е. $w_n < W < w_v$ нет оснований отвергать H_0 об однородности выборок.

Если требуется проверить на уровне значимости β=2Q нулевую гипотезу, состоящую в том, что при всех значениях аргумента функции распределения равны между собой, то есть H_0: $F_1(x)$=$F_2(x)$ при конкурирующей гипотезе H_1: $F_1(x) > F_2(x)$, то при определении нижней критической точки нужно учитывать что Q=β.

Если $W > w_n$ – нет оснований отвергать H_0; если $W < w_v$ – гипотезу отвергают.

Если конкурирующая гипотеза H_1 задана условием $F_1(x) < F_2(x)$, то верхняя критическая точка находится по формуле

$$w_v=(n_1+n_2+1)n_1-w_n(Q;n_1;n_2),$$

где Q=β.

При $W < w_v$ – нет оснований отвергать H_0; при $W > w_n$ – нулевую гипотезу отвергают.

Замечание. Если несколько вариант только одной выборки одинаковы, то в общем вариационном ряду им присваивают обычные порядковые номера (совпадающие варианты нумеруют так, как если бы они были различными числами); если же совпадают варианты разных выборок, то всем им присваивают один и тот же порядковый номер, равный среднему арифметическому порядковых номеров, которые имели бы эти варианты до совпадения.

Точечные и интервальные оценки параметров распределения

Известно, моделирование проводится для определения тех или иных характеристик системы (например, качества системы обнаружения полезного сигнала в помехах, измерения ее параметров). Так как входные воздействия представляют собой случайные процессы, то для получения информации о параметрах системы необходимо провести определенное количество (как правило, достаточно большое) экспериментов или опытов.

Важной задачей моделирования систем является задача *оценивания* (приближенного определения) по выборочным данным параметров закона распределения признака X генеральной совокупности. Другими словами, необходимо по данным *выборочного распределения* оценить неизвестные параметры *теоретического распределения*. Статистические *оценки* могут быть *точечными* и *интервальными*.

Точечной оценкой неизвестного параметра называют *число* (точку на числовой оси), которое *приблизительно равно* оцениваемому параметру и может *заменить* его *с достаточной степенью точности* в статистических расчетах.

Для того чтобы точечные статистические оценки обеспечивали "хорошие" приближения неизвестных параметров, они должны быть *несмещенными*, *состоятельными* и *эффективными.*

Пусть θ^* – точечная оценка неизвестного параметра θ.

Несмещенной называют такую точечную статистическую *оценку* θ^*, *математическое ожидание* которой *равно* оцениваемому *параметру*: $M\left(\theta^*\right)=\theta$.

Состоятельной называют такую точечную статистическую *оценку*, которая при $n \to \infty$ стремится по вероятности к оцениваемому *параметру*. В частности, если дисперсия несмещенной оценки при $n \to \infty$ стремится к нулю, то такая оценка оказывается и состоятельной.

Эффективной называют такую точечную статистическую *оценку*, которая при фиксированном n имеет *наименьшую дисперсию*.

Оценка математического ожидания

Пусть имеется случайная величина X с математическим ожиданием m и дисперсией D, при этом оба эти параметра неизвестны. Над величиной X произведено N независимых экспериментов, в результате которых была получена совокупность N численных результатов x_1, x_2, …, x_N. В качестве оценки математического ожидания естественно предложить среднее арифметическое наблюдаемых значений

$$\grave{o}^* = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

Оценка математического ожидания является несмещенной.

Оценка дисперсии

При больших объемах выборки для оценки дисперсии используют формулу:

$$D^* = \frac{\sum_{i=1}^{N}(x_i - m^*)^2}{N}.$$

При относительно малых выборках следует пользоваться формулой для исправленной дисперсии:

$$S^2 = D^{**} = \frac{D^* N}{N-1}.$$

Оценку среднего квадратического отклонения (стандарта) производят по формуле:

$$S = \sqrt{S^2}\ .$$

Интервальные оценки

Интервальной называют оценку, которая определяется двумя числами – концами отрезка.

Для построения *интервальной оценки* рассмотрим событие, заключающееся в том, что отклонение точечной оценки параметра θ^* от истинного значения этого параметра θ по абсолютной величине не превышает некоторую положительную величину Δ. Вероятность такого события $P\left(\left|\theta - \theta^*\right| < \Delta\right) = \gamma$. Заменив неравенство $\left|\theta - \theta^*\right| < \Delta$ на равносильное, получим:

$$P\left(\theta^* - \Delta < \theta < \theta^* + \Delta\right) = \gamma\ .$$

Вероятность того, что *доверительный интервал* $(\theta^* - \Delta,\ \theta^* + \Delta)$ заключает в себе (покрывает) неизвестный параметр θ равна γ и называется *доверительной вероятностью* или *надежностью* интервальной оценки. Величину Δ называют *точностью* оценки.

Иногда при выборе интервала за аргумент вместо доверительной вероятности γ берут величину, дополняющую ее до 1, т.е. $1-\gamma$ (вероятность выхода за интервал). Эту величину называют *уровнем значимости* β.

I. Построим интервальную оценку математического ожидания для двух случаев:

1) параметр σ признака X *известен*. В этом случае *интервальная оценка* параметра a с заданной *надежностью* γ определяется формулой:

$$\bar{x}_{\text{в}} - \Delta < a < \bar{x}_{\text{в}} + \Delta,$$

где $\Delta = \dfrac{t \cdot \sigma}{\sqrt{n}}$, t – аргумент функции Лапласа: $\Phi(t) = \dfrac{\gamma}{2}$.

2) параметр σ признака X *неизвестен*. В этом случае *интервальная оценка* параметра a с заданной *надежностью* γ определяется формулой:

$$\overline{x}_в - \Delta < a < \overline{x}_в + \Delta ,$$

где $\Delta = \frac{t_\gamma \cdot S}{\sqrt{n}}$, S – точечная оценка параметра σ, $t_\gamma = t(\gamma, n)$ – значения распределения Стьюдента.

II. Построим интервальную оценку квадратического отклонения.

Интервальной оценкой среднего квадратического отклонения признака X служит доверительный интервал:

$$s(1-q) < \sigma < s(1+q) \text{ (при } q < 1),$$

$$0 < \sigma < s(1+q) \text{ (при } q > 1).$$

Для отыскания q пользуются таблицей значений $q=q(\gamma,n)$.

Пример 1. На станции технического обслуживания автомобилей исследовались затраты времени на ремонт карбюратора. Были зафиксированы следующие результаты:

Затраты времени (в мин.)	22–24	24–26	26–28	28–30	30–32	32–34
Число наблюдений, попавших в данный интервал	2	12	34	40	10	2

Построить доверительные интервалы для a и σ при доверительной вероятности γ=0,95.

Решение.

Признак X – затраты времени на ремонт карбюратора (мин) – непрерывный. Распределение задано интервальным рядом. Характеристики такого ряда находят по тем же формулам, что и для дискретного ряда, предварительно заменив интервальный ряд дискретным. Для этого каждый интервал x_{i-1}–x_i заменяется его серединой x_i'. Расчеты представим в таблице:

Затраты времени (X, мин): x_{i-1}–x_i	Число наблюдений (m_i)	x_i'	$x_i' m_i$	$(x_i')^2 m_i$
22–24	2	23	46	1058
24–26	12	25	300	7500
26–28	34	27	918	24786
28–30	40	29	1160	33640
30–32	10	31	310	9610
32–34	2	33	66	2178
Итого	100	-	2800	78772

$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i' \cdot m_i = \frac{2800}{100} =$ 28 (мин) – среднее время на обработку одной детали.

Дисперсию рассчитываем по формуле:

$$D_{\text{в}} = \overline{\left(x'\right)^2} - \overline{x}_{\text{в}}^2,$$

где

$$\overline{\left(x'\right)^2} = \frac{1}{n}\sum_{i=1}^{k}\left(x_i'\right)^2 \cdot m_i = \frac{78772}{100} = 787,72;$$

$$D_{\hat{a}} = 787,72 - (28)^2 = 3,72;$$

$$\sigma_{\hat{a}} = \sqrt{D_{\hat{a}}} = \sqrt{3,72} \approx 1,93 \text{ мин.}$$

Параметр σ признака X известен. В этом случае интервальная оценка параметра a с заданной надежностью γ определяется формулой:

$$\overline{x}_{\hat{a}} - \Delta < a < \overline{x}_{\hat{a}} + \Delta,$$

где $\Delta = \dfrac{t \cdot \sigma}{\sqrt{n}}$, t – аргумент функции Лапласа: $\Phi(t) = \dfrac{\gamma}{2}$.

$\Phi(t) = \dfrac{\gamma}{2} = 0,475$. Тогда t=1,96; $\Delta = \dfrac{t \cdot \sigma}{\sqrt{n}} = \dfrac{1,96 \cdot 1,93}{\sqrt{100}} = 0,38$.

Следовательно, интервальная оценка параметра a с заданной надежностью γ имеет вид:

$$28 - 0,38 < a < 28 + 0,38;$$

$$27,62 < a < 28,38.$$

По таблице значений $q=q(\gamma,n)= q(0,95;100)=0,143$.

Искомый доверительный интервал таков:

$$s(1-q) < \sigma < s(1+q);$$

$$1,93(1-0,143) < \sigma < 1,93(1+0,143);$$

$$1,65 < \sigma < 2,2.$$

Оценка вероятности по частоте появления события

При проведении экспериментов часто приходится оценивать неизвестную вероятность события P по его частоте P^* в N независимых экспериментах. Частота некоторого события в N независимых экспериментах есть не что иное, как среднее арифметическое наблюдаемых значений величины X, которая в каждом отдельном опыте принимает значение 1 (если событие совершилось), или значение 0 (если событие не произошло):

$$P^* = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Интервальной оценкой вероятности по частоте появления события P^* в N независимых экспериментах служит доверительный интервал

$$I=(P_1,P_2),$$

где

$$P_1 = \frac{N}{t^2+N}[P^* + \frac{t^2}{2N} - t\sqrt{\frac{P^*(1-P^*)}{N} + \left(\frac{t}{2N}\right)^2}],$$
$$P_2 = \frac{N}{t^2+N}[P^* + \frac{t^2}{2N} + t\sqrt{\frac{P^*(1-P^*)}{N} + \left(\frac{t}{2N}\right)^2}]. \qquad (1)$$

Замечание. При больших значениях N (порядка сотен) можно принять в качестве приближенных границ доверительного интервала:

$$P_1 = P^* - t\sqrt{\frac{P^*(1-P^*)}{N}},$$
$$P_2 = P^* + t\sqrt{\frac{P^*(1-P^*)}{N}}. \qquad (2)$$

Здесь P^* – конкретная оценка вероятности (частоты события), а t находится, исходя из заданной доверительной вероятности γ:

$$\Phi(t) = \frac{\gamma}{2}.$$

Пример 2. Производится серия из 60 экспериментов с целью оценки вероятности некоторого события. В результате этой серии экспериментов событие A появилось 15 раз. Построить доверительный интервал, в котором с вероятностью 0,95 вычисляется истинная вероятность рассматриваемого события.

Решение: По условию, N=60, m=15, γ=0,95. Найдем относительную частоту появления события A – P^*. $P^*=m/N$=15/60=0,25.

По таблице значений функции $\Phi(x)$ при γ=0,95 находим: t = 1,96.

Подставив в формулу (1) полученные значения, получим P_1=0,16, P_2=0,37.

Итак, искомый доверительный интервал (0,16;0,37).

Пример 3. Производится серия из 200 экспериментов с целью оценки вероятности некоторого события. В результате этой серии экспериментов получено значение P^* = 0,34. Построить доверительный интервал, в котором с вероятностью 0,85 вычисляется истинная вероятность рассматриваемого события.

Решение: По таблице значений функции $\Phi(x)$ при γ=0,85 находим: t=1,439. Умножая это значение на величину

$$\sqrt{\frac{P^*(1-P^*)}{N}} \approx 0{,}0335,$$

получим 0,048. Откуда искомый доверительный интервал

$$I \approx (0{,}292; 0{,}388).$$

Рекуррентные соотношения для дисперсии и математического ожидания

При моделировании удобно пользоваться рекуррентной формулой для оценки математического ожидания:

$$m_N^* = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{N-1}{N} m_{N-1}^* + \frac{1}{N} x_N ,$$

где m_N^* – оценка математического ожидания по выборке x_i из N отсчетов, m_{N-1}^* – оценка математического ожидания по выборке x_i из (N–1) отсчета.

Аналогичная формула используется для рекуррентной оценки дисперсии:

$$D_N^* = [(N-1)/N]D_{N-1}^* + [(N-1)/N^2][x_N - m_{N-1}^*]^2,$$

где D_N^* – оценка дисперсии по выборке x_i из N отсчетов, D_{N-1}^* – оценка дисперсии по выборке x_i из (N–1) отсчета.

Пример 4. В результате эксперимента получена следующая реализация: 5, 1, 6, 2, 12. Построить рекуррентную оценку математического ожидания и дисперсии.

Решение: При построении рекуррентной оценки математического ожидания воспользуемся формулой

$$m_N^* = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{N-1}{N} m_{N-1}^* + \frac{1}{N} x_N .$$

Тогда получим:

$$m_1^* = x_1 = 5,$$

$$m_2^* = 1/2 m_1^* + 1/2 x_2,\ m_2^* = 3,$$

$$m_3^* = 2/3 m_2^* + 1/3 x_3,\ m_3^* = 4,$$

$$m_4^* = 3/4 m_3^* + 1/4 x_4,\ m_4^* = 3{,}5,$$

$$m_5^* = 4/5 m_4^* + 1/5 x_5,\ m_5^* = 5{,}2.$$

Таким образом, оценка математического ожидания равна $m^* = 5{,}2$.

Найдем оценку математического ожидания по формуле

$$m^* = \frac{1}{N}\sum_{i=1}^{N} X_i ,$$

$$m^* = 1/5 \cdot (5+1+6+2+12) = 5{,}2.$$

Для построения рекуррентной оценки дисперсии воспользуемся формулой

$$D_N^* = [(N-1)/N]D_{N-1}^* + [(N-1)/N^2][x_N - m_{N-1}^*]^2.$$

Таким образом, имеем:

$$D_1^* = 0,$$

$$D_2^* = 1/2D_1^* + 1/4(x_2 - m_1^*)^2,\ D_2^* = 4,$$

$$D_3^* = 2/3D_2^* + 2/9(x_3 - m_2^*)^2,\ D_3^* = 4{,}67,$$

$$D_4^* = 3/4D_3^* + 3/16(x_4 - m_3^*)^2,\ D_4^* = 4{,}25,$$

$$D_5^* = 4/5D_4^* + 4/25(x_5 - m_4^*)^2,\ D_5^* = 14{,}96.$$

Получили, что оценка дисперсии равна D*=14,96.

Найдем оценку дисперсии по формуле

$$D^* = \frac{\sum_{i=1}^{N}(X_i - m^*)^2}{N}.$$

D*=1/5[(5–5,2)^2+(1–5,2)^2+(6–5,2)^2+(2–5,2)^2+(12–5,2)^2]=14,96.

Статистическая проверка гипотез

С теорией статистического оценивания параметров тесно связана проверка статистических гипотез. Она используется всякий раз, когда необходим обоснованный вывод о преимуществах того или иного способа инвестиций, измерений, стрельбы, технологического прогресса, об эффективности нового метода обучения, управления, о пользе вносимого удобрения, лекарства, об уровне доходности ценных бумаг, о значимости математической модели и т.д.

Статистической гипотезой называется любое предположение о виде или параметрах неизвестного закона распределения.

Различают *простую* и *сложную* статистические гипотезы. При простых гипотезах обязательно известен закон распределения исследуемой случайной величины и оценка математического ожидания строго равна табличному значению. При проверке сложных гипотез не известен закон распределения случайной величины, и нет строгого равенства между табличным значением и оценкой математического ожидания.

Проверяемую гипотезу обычно называют *нулевой* (или основной) и обозначают H_0 . Наряду с нулевой гипотезой H_0 рассматривают *альтернативную*, или *конкурирующую*, гипотезу H_1, являющуюся логическим отрицанием H_0. Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого в задачах проверки статистических гипотез. Правило, по которому гипотеза H_0 отвергается или принимается, называется *статистическим критерием* или *статистическим тестом*.

Все проверки гипотез дают ответ с определенной степенью надежности. Для этого задается уровень значимости.

Вероятность α допустить ошибку 1-го рода, то есть отвергнуть гипотезу H_0, когда она верна, называется *уровнем значимости критерия*. Вероятность допустить ошибку 2-го рода, то есть принять гипотезу H_0, когда она неверна, обычно обозначают β.

Вероятность $(1-\beta)$ не допустить ошибку 2-го рода, то есть отвергнуть гипотезу H_0, когда она неверна, называется *мощностью (или функцией мощности) критерия*.

Наблюдаемым (эмпирическим) значением $K_{набл}$ называют то значение критерия, которое вычислено по выборке.

Для проверки гипотез вводится случайная величина, для которой обязательно известен закон распределения. Эта случайная величина практически реализуется на интервале, который мы будем разделять на зону принятия и не принятия гипотезы. Чтобы указать эту зону не принятия гипотезы, задаем уровень значимости α.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (область допустимых значений) называют совокупность значений критерия, при которых нулевую гипотезу принимают.

Критическими точками (границами) $K_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Правосторонней называют критическую область, определяемую неравенством $K > K_{кр}$, где $K_{кр} > 0$.

Левосторонней называют критическую область, определяемую неравенством $K < K_{кр}$, где $K_{кр} < 0$.

Двусторонней называют критическую область, определяемую неравенством $K < K_1$, $K > K_2$. В частности, если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенством $K < -K_{кр}$, $K > K_{кр}$ $(K_{кр} > 0)$.

Для отыскания критической области задаются уровнем значимости α и ищут критические точки, исходя из следующих соотношений:

а) для правосторонней критической области $P(K > K_{кр}) = \alpha$, $(K_{кр} > 0)$;

б) для левосторонней критической области $P(K < K_{кр}) = \alpha$, $(K_{кр} < 0)$;

в) для двусторонней критической области:

$$\begin{cases} P(K > K_{кр}) = \dfrac{\alpha}{2}, \\ P(K < -K_{кр}) = \dfrac{\alpha}{2}. \end{cases}$$

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

Приложение 1

Таблица значений функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,32	0,1255	0,64	0,2389	0,96	0,3315
0,01	0,0040	0,33	0,1293	0,65	0,2422	0,97	0,3340
0,02	0,0080	0,34	0,1331	0,66	0,2454	0,98	0,3365
0,03	0,0120	0,35	0,1368	0,67	0,2486	0,99	0,3389
0,04	0,0160	0,36	0,1406	0,68	0,2516	1,00	0,3413
0,05	0,0199	0,37	0,1443	0,69	0,2549	1,01	0,3438
0,06	0,0239	0,38	0,1480	0,70	0,2580	1,02	0,3461
0,07	0,0279	0,39	0,1517	0,71	0,2611	1,03	0,3485
0,08	0,0319	0,40	0,1554	0,72	0,2642	1,04	0,3508
0,09	0,0359	0,41	0,1591	0,73	0,2673	1,05	0,3531
0,10	0,0398	0,42	0,1628	0,74	0,2703	1,06	0,3554
0,11	0,0438	0,43	0,1664	0,75	0,2734	1,07	0,3577
0,12	0,0478	0,44	0,1700	0,76	0,2764	1,08	0,3599
0,13	0,0517	0,45	0,1736	0,77	0,2794	1,09	0,3621
0,14	0,0557	0,46	0,1772	0,78	0,2823	1,10	0,3643
0,15	0,0596	0,47	0,1808	0,79	0,2852	1,11	0,3665
0,16	0,0636	0,48	0,1844	0,80	0,2881	1,12	0,3686
0,17	0,0675	0,49	0,1879	0,81	0,2910	1,13	0,3708
0,18	0,0714	0,50	0,1915	0,82	0,2939	1,14	0,3729
0,19	0,0753	0,51	0,1950	0,83	0,2967	1,15	0,3749
0,20	0,0793	0,52	0,1985	0,84	0,2995	1,16	0,3770
0,21	0,0832	0,53	0,2019	0,85	0,3023	1,17	0,3790
0,22	0,0871	0,54	0,2054	0,86	0,3051	1,18	0,3810
0,23	0,0910	0,55	0,2088	0,87	0,3078	1,19	0,3830
0,24	0,0948	0,56	0,2123	0,88	0,3106	1,20	0,3849
0,25	0,0987	0,57	0,2157	0,89	0,3133	1,21	0,3869
0,26	0,1026	0,58	0,2190	0,90	0,3159	1,22	0,3883
0,27	0,1064	0,59	0,2224	0,91	0,3186	1,23	0,3907
0,28	0,1103	0,60	0,2257	0,92	0,3212	1,24	0,3925
0,29	0,1141	0,61	0,2291	0,93	0,3228	1,25	0,3944
0,30	0,1179	0,62	0,2324	0,94	0,3264	1,26	0,3962
0,31	0,1217	0,63	0,2357	0,95	0,3289	1,27	0,3980

Окончание прил.1

x	$Ф(x)$	x	$Ф(x)$	x	$Ф(x)$	x	$Ф(x)$
1,28	0,3997	1,61	0,4463	1,94	0,4738	2,54	0,4945
1,29	0,4015	1,62	0,4474	1,95	0,4744	2,56	0,4948
1,30	0,4032	1,63	0,4484	1,96	0,4750	2,58	0,4951
1,31	0,4049	1,64	0,4495	1,97	0,4756	2,60	0,4953
1,32	0,4066	1,65	0,4505	1,98	0,4761	2,62	0,4956
1,33	0,4082	1,66	0,4515	1,99	0,4767	2,64	0,4959
1,34	0,4099	1,67	0,4525	2,00	0,4772	2,66	0,4961
1,35	0,4115	1,68	0,4535	2,02	0,4783	2,68	0,4963
1,36	0,4131	1,69	0,4545	2,04	0,4793	2,70	0,4965
1,37	0,4147	1,70	0,4554	2,06	0,4803	2,72	0,4967
1,38	0,4162	1,71	0,4564	2,08	0,4812	2,74	0,4969
1,39	0,4177	1,72	0,4573	2,10	0,4821	2,76	0,4971
1,40	0,4192	1,73	0,4582	2,12	0,4830	2,78	0,4973
1,41	0,4207	1,74	0,4591	2,14	0,4838	2,80	0,4974
1,43	0,4236	1,76	0,4608	2,18	0,4854	2,84	0,4977
1,44	0,4251	1,77	0,4616	2,20	0,4861	2,86	0,4979
1,45	0,4265	1,78	0,4625	2,22	0,4868	2,88	0,4980
1,46	0,4279	1,79	0,4633	2,24	0,4875	2,90	0,4981
1,47	0,4292	1,80	0,4641	2,26	0,4881	2,92	0,4982
1,48	0,4306	1,81	0,4649	2,28	0,4887	2,94	0,4984
1,49	0,4319	1,82	0,4656	2,30	0,4893	2,96	0,4985
1,50	0,4332	1,83	0,4664	2,32	0,4898	2,98	0,4986
1,51	0,4345	1,84	0,4671	2,34	0,4904	3,00	0,49865
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,49931
1,53	0,4370	1,86	0,4686	2,38	0,4913	3,40	0,49966
1,54	0,4382	1,87	0,4693	2,40	0,4918	3,60	0,499841
1,55	0,4394	1,88	0,4699	2,42	0,4922	3,80	0,499928
1,56	0,4406	1,89	0,4706	2,44	0,4927	4,00	0,499968
1,57	0,4418	1,90	0,4713	2,46	0,4931	4,50	0,499997
1,58	0,4429	1,91	0,4719	2,48	0,4934	5,00	0,499997
1,59	0,4441	1,92	0,4726	2,50	0,4938		
1,60	0,4452	1,93	0,4732	2,52	0,4941		

Приложение 2

Квантили распределения χ^2

Число степеней свободы	Уровень значимости χ^2					
k	0,01	0,025	0,05	0,95	0,975	0, 99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	о,аз1	0,554
6	16,8	14,4	12,6	1.64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,50
Н	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5.23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11.7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
25	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16.0	14,3
30	50$_{9}$9	47,0	43,8	18,5	16,8	15,0

Приложение 3

Критические значения для наибольшего отклонения эмпирического распределения от теоретического

Объем выборки N	Доверительная вероятность α			
	0,9	0,95	0,98	0,999
	Уровень значимости $\beta, \%$			
	10	5	2	1
1	0,9500	0,9750	0,9900	0,9950
2	0,7764	0,8419	0,9000	0,9293
3	0,6360	0,7076	0,7846	0,8290
4	0,5652	0,6239	0,6889	0,7342
5	0,5095	0,5633	0,6272	0,6685
6	0,4680	0,5193	0,5774	0,6167
7	0,4361	0,4834	0,5384	0,5758
8	0,4096	0,4543	0,5064	0,5418
9	0,3875	0,4300	0,4796	0,5133
10	0,3687	0,4093	0,4566	0,4889
11	0,3524	0,3912	0,4367	0,4677
12	0,3382	0,3754	0,4192	0,4491
13	0,3255	0,3614	0,4036	0,4325
14	0,3142	0,3489	0,3897	0,4176
15	0,3040	0,3376	0,3771	0,4042
16	0,2947	0,3273	0,3657	0,3920
17	0,863	0,3180	0,3553	0,3809
18	0,2785	0,3094	0,3457	0,3706
19	0,2714	0,3014	0,3369	0,3612
20	0,2647	0,2941	0,3287	0,3524
25	0,2377	0,2644)	0,2952	0,3166
30	0,2176	0,2417	0,2702	0,2899
35	0,2019	0,2243	0,2507	0,2690
> 35	$1{,}21/\sqrt{N}$	$1{,}34/\sqrt{N}$	$1{,}50/\sqrt{N}$	$1{,}61/\sqrt{N}$

Приложение 4

Распределение Вилкоксона

Объемы выборки		$\alpha/2$			Объемы выборки		$\alpha/2$		
n_1	n_2	0,05	0,025	0,005	n_1	n_2	0,05	0,025	0,005
7	7	39	36	32	8	8	51	49	43
	8	41	38	34		9	54	51	45
	9	43	40	35		10	56	53	47
	10	45	42	37		11	59	55	49
	11	47	44	38		12	62	58	51
	12	49	46	40		13	64	60	53
	13	52	48	41		14	67	62	54
	14	54	50	43		15	69	65	56
	15	56	52	44		16	72	67	58
9	9	66	62	56	10	10	82	78	71
	10	69	65	58		11	86	81	73
	11	72	68	61		12	89	84	76
	12	75	71	63		13	92	88	79
11	11	100	96	87	12	11	100	96	87
	12	104	99	90		12	104	99	90
	13	108	103	93		13	108	103	93
	14	112	106	96		14	112	106	96
	15	116	110	99		15	116	110	99