Владимир Гмурман. Теория вероятностей и математическая статистика

Многие поколения студентов хорошо знают это пособие, ставшее классическим учебным изданием. Его ценность заключается в том, что сложные вопросы теории вероятностей и математической статистики изложены в логической последовательности и доступной форме. Большое количество примеров позволяет лучше усвоить материал, а задачи, приведенные в конце каждой главы, – закрепить полученные знания. Мои комментарии даны с отступом.

Владимир Гмурман. Теория вероятностей и математическая статистика. – М.: Издательство Юрайт, 2023. – 480 с.

Купить цифровую книгу в ЛитРес, бумажную книгу в Ozon

ЧАСТЬ ПЕРВАЯ. СЛУЧАЙНЫЕ СОБЫТИЯ

Глава первая. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Классическое определение. Вероятностью события А называют отношение числа благоприятствующих этому событию исходов к общему числу всех равновозможных несовместных элементарных исходов, образующих полную группу.

Глава вторая. ТЕОРЕМА СЛОЖЕНИЯ ВЕРОЯТНОСТЕЙ

Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий:

$$(2.1)\ P(A+B) = P(A) + P(B)$$

Сумма вероятностей событий $A_1, A_2, \dots A_n$, образующих полную группу, равна единице:

$$(2.2)\ P(A_1) + P(A_2) + \cdots + P(A_n) = 1$$

Сумма вероятностей противоположных событий равна единице:

$$(2.3)\ P(A) + P(\bar{A}) = 1$$

Глава третья. ТЕОРЕМА УМНОЖЕНИЯ ВЕРОЯТНОСТЕЙ

Произведением двух событий А и В называют событие АВ, состоящее в совместном появлении (совмещении) этих событий.

Условной вероятностью $P_A(B)$ называют вероятность события В, вычисленную в предположении, что событие А уже наступило.

Вероятность совместного появления двух событий равна произведению вероятности одного из них на условную вероятность другого, вычисленную в предположении, что первое событие уже наступило:

$$(3.1)\ P(AB) = P(A)P_A(B)$$

Событие B называют независимым от события A, если появление события A не изменяет вероятности события B, т. е. если условная вероятность события В равна его безусловной вероятности:

$$(3.2)\ P_A(B) = P(B)$$

Для независимых событий теорема умножения (3.1) имеет вид

$$(3.3)\ P(AB) = P(A)P(B)$$

Вероятность появления хотя бы одного из событий $A_1, A_2, \dots A_n$, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий $\bar{A}_1, \bar{A}_2, \dots \bar{A}_n$:

$$(3.4)\ P(AB) = 1 - q_1 q_2 \dots q_n$$

Если события $A_1, A_2, \dots A_n$ имеют одинаковую вероятность, равную p, то вероятность появления хотя бы одного из этих событий

$$(3.5)\ P(AB) = 1 - q^n$$

Глава четвертая. СЛЕДСТВИЯ ТЕОРЕМ СЛОЖЕНИЯ И УМНОЖЕНИЯ

Вероятность появления хотя бы одного из двух совместных событий равна сумме вероятностей этих событий без вероятности их совместного появления:

$$(4.1)\ P(A + B) = P(A) + P(B) - P(AB)$$

Если события A и B несовместны, то их совмещение есть невозможное событие и, следовательно, $P(AB)$ = 0. Формула (4.1) для несовместных событий принимает вид

$$(4.2)\ P(A + B) = P(A) + P(B)$$

Вероятность события A, которое может наступить лишь при условии появления одного из несовместных событий $B_1, B_2, \dots, B_n$, образующих полную группу, равна сумме произведений вероятностей каждого из этих событий на соответствующую условную вероятность события A:

$$(4.3)\ P(A) = P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A) + \dots + P(B_n)P_{B_n}(A)$$

Вероятность гипотез. Формулы Байеса

Пусть событие A может наступить при условии появления одного из несовместных событий $B_1, B_2, \dots, B_n$, образующих полную группу. Поскольку заранее не известно, какое из этих событий наступит, их называют гипотезами. Вероятность появления события А определяется формуле полной вероятности (4.3).

Допустим, что произведено испытание, в результате которого появилось событие A. Поставим своей задачей определить, как изменились (в связи с тем, что событие A уже наступило) вероятности гипотез. Другими словами, будем искать условные вероятности $P_A(B_1), P_A(B_2), \dots, P_A(B_n)$.

Найдем сначала условную вероятность $P_A(B_1)$. По теореме умножения имеем

$$(4.4)\ P(AB_1) = P(A)P_A(B_1) = P(B_1)P_{B_1}(A)$$

Отсюда

$$(4.5)\ P_A(B_1) = \frac{P(B_1)P_{B_1}(A)}{P(A)}$$

Заменив здесь $P(A)$ по формуле (4.3), получим

$$(4.6)\ P_A(B_1) = \frac{P(B_1)P_{B_1}(A)}{P(B_1)P_{B_1}(A) + P(B_2)P_{B_2}(A) + \dots + P(B_n)P_{B_n}(A)}$$

Формулы (4.5) и (4.6) называют формулами Бейеса (по имени английского математика, который их вывел; опубликованы в 1764 г.). Формулы Бейеса позволяют переоценить вероятности гипотез после того, как становится известным результат испытания, в итоге которого появилось событие A.

Глава пятая. ПОВТОРЕНИЕ ИСПЫТАНИЙ

Формула Бернулли

Пусть производится n независимых испытаний, в каждом из которых событие A может появиться либо не появиться. Условимся считать, что вероятность события A в каждом испытании одна и та же –

p. Следовательно, вероятность ненаступления события A в каждом испытании также постоянна и равна $q = 1 - p$.

Вероятность того, что при n испытаниях событие A осуществится ровно k раз определяется формулой Бернулли:

$$(5.1)\ P_n(k) = \frac{n!}{k!\,(n-k)!} p^k q^{n-k} = C_n^k p^k q^{n-k}$$

Локальная теорема Лапласа дает асимптотическую формулу, которая позволяет приближенно найти вероятность появления события ровно k раз в n испытаниях, если число испытаний достаточно велико. Для частного случая $p = 1/2$ асимптотическая формула была найдена в 1730 г. Муавром; в 1783 г. Лаплас обобщил формулу Муавра для произвольного p, отличного от 0 и 1. Поэтому теорему иногда называют теоремой Муавра – Лапласа.

Если вероятность p появления события A в каждом испытании постоянна и отлична от нуля и единицы, то вероятность $P_n(k)$ того, что событие A появится в n испытаниях ровно k раз, приближенно равна (тем точнее, чем больше n) значению функции

$$(5.2)\ P_n(k) \approx \frac{1}{\sqrt{npq}}\varphi(x), \qquad \text{где } \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad x = \frac{k-np}{\sqrt{npq}}$$

Имеются таблицы, в которых помещены значения функции $\varphi(x)$ соответствующие положительным значениям аргумента x (рис. 1). Для отрицательных значений аргумента пользуются теми же таблицами, так как функция $\varphi(x)$ четна, т. е. $\varphi(-x) = \varphi(x)$.

	0	1	2	3	4	5	6	7	8	9
0	0,39894	0,39892	0,39886	0,39876	0,39862	0,39844	0,39822	0,39797	0,39767	0,39733
0,1	0,39695	0,39654	0,39608	0,39559	0,39505	0,39448	0,39387	0,39322	0,39253	0,39181
0,2	0,39104	0,39024	0,38940	0,38853	0,38762	0,38667	0,38568	0,38466	0,38361	0,38251
0,3	0,38139	0,38023	0,37903	0,37780	0,37654	0,37524	0,37391	0,37255	0,37115	0,36973
0,4	0,36827	0,36678	0,36526	0,36371	0,362					0,35381
0,5	0,35207	0,35029	0,34849	0,34667	0,344					0,33521
0,6	0,33322	0,33121	0,32918	0,32713	0,325					0,31443
0,7	0,31225	0,31006	0,30785	0,30563	0,303					0,29200
0,8	0,28969	0,28737	0,28504	0,28269	0,280					0,26848
0,9	0,26609	0,26369	0,26129	0,25888	0,256					0,24439
1	0,24197	0,23955	0,23713	0,23471	0,232					0,22025
1,1	0,21785	0,21546	0,21307	0,21069	0,208					0,19652
1,2	0,19419	0,19186	0,18954	0,18724	0,184					0,17360
1,3	0,17137	0,16915	0,16694	0,16474	0,162					0,15183
1,4	0,14973	0,14764	0,14556	0,14350	0,141					0,13147
1,5	0,12952	0,12758	0,12566	0,12376	0,121					0,11270
1,6	0,11092	0,10915	0,10741	0,10567	0,10396	0,10226	0,10059	0,09893	0,09728	0,09566

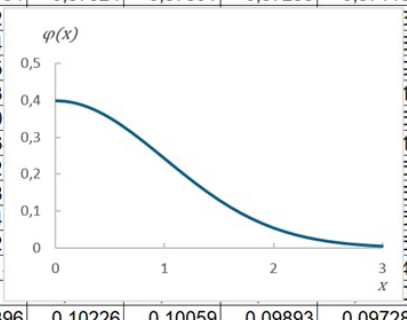

Рис. 1. Таблица значений функции Лапласа $\varphi(x)$

В Excel функции Лапласа $\varphi(x)$ соответствует функция =НОРМ.СТ.РАСП(x;ЛОЖЬ).

Интегральная теорема Лапласа

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

$$(5.3)\ P_n(k_1,k_2) \simeq \frac{1}{\sqrt{2\pi}} \int_{x'}^{x''} e^{-\frac{z^2}{2}}\,dz, \qquad \text{где } x' = \frac{k_1-np}{\sqrt{npq}}, x'' = \frac{k_2-np}{\sqrt{npq}}$$

При решении задач, требующих применения интегральной теоремы Лапласа, пользуются таблицами, так как неопределенный интеграл не выражается через элементарные функции. Таблица для интеграла

$$(5.4)\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}}\,dz$$

	0	1	2	3	4	5	6	7	8	9
0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,170					0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,205					0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,238					0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,270					0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,299					0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,326					0,33891
1	0,34134	0,34375	0,34614	0,34849	0,350					0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,372					0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,392					0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,409					0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,425					0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,438					0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327

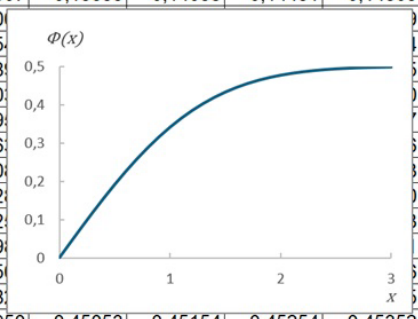

Рис. 2. Таблица значений интегральной функции Лапласа $\Phi(x)$

В Excel интегральную функцию Лапласа $\Phi(x)$ можно найти с помощью формулы =НОРМ.СТ.РАСП(x;ИСТИНА) – 0,5.

ЧАСТЬ ВТОРАЯ. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

Глава шестая. ВИДЫ СЛУЧАЙНЫХ ВЕЛИЧИН. ЗАДАНИЕ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Случайной называют величину, которая в результате испытания примет одно и только одно возможное значение, наперед не известное и зависящее от случайных причин, которые заранее не могут быть учтены.

Законом распределения дискретной случайной величины называют соответствие между возможными значениями и их вероятностями; его можно задать таблично, аналитически (в виде формулы) и графически.

При табличном задании закона распределения дискретной случайной величины первая строка таблицы содержит возможные значения, а вторая — их вероятности:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Биномиальное распределение

Пусть производится n независимых испытаний, в каждом из которых событие A может появиться либо не появиться. Вероятность наступления события во всех испытаниях постоянна и равна p (следовательно, вероятность непоявления $q = 1 - p$). Рассмотрим в качестве дискретной случайной величины X число появлений события A в этих испытаниях.

Поставим перед собой задачу: найти закон распределения величины X. Для ее решения требуется определить возможные значения X и их вероятности. Очевидно, событие A в n испытаниях может либо не появиться, либо появиться 1 раз, либо 2 раза, ..., либо n раз. Таким образом, возможные значения X таковы: $x_1 = 0, x_2 = 1, x_3 = 2, \ldots, x_{n+1} = n$. Остается найти вероятности этих возможных значений, для чего достаточно воспользоваться формулой Бернулли (2.15). Эта формула является аналитическим выражением искомого закона распределения.

Биномиальным называют распределение вероятностей, определяемое формулой Бернулли. Закон назван «биномиальным» потому, что правую часть равенства (2.15) можно рассматривать как общий член разложения бинома Ньютона:

$$(6.1)\ (p+q)^n = C_n^n p^n + C_n^{n-1} p^{n-1} q + \cdots + C_n^k p^k q^{n-k} + \cdots + C_n^0 q^n$$

Таким образом, первый член разложения p^n определяет вероятность наступления рассматриваемого события n раз в n независимых испытаниях; второй член $np^{n-1}q$ определяет вероятность наступления события $n - 1$ раз; ...; последний член q^n определяет вероятность того, что событие не появится ни разу.

Напишем биномиальный закон в виде таблицы:

X	n	$n-1$	…	k	…	0
P	np^n	$np^{n-1}q$	…	$C_n^k p^k q^{n-k}$	…	q^n

Распределение Пуассона

Пусть производится n независимых испытаний, в каждом из которых вероятность появления события A равна p. Для определения вероятности k появлений события в этих испытаниях используют формулу Бернулли. Если же n велико, то пользуются асимптотической формулой Лапласа. Однако эта формула непригодна, если вероятность события мала ($p \leq 0{,}1$). В этих случаях (n велико, p мало) прибегают к асимптотической формуле Пуассона.

Если произведение np сохраняет постоянное значение, а именно $np = \lambda$, то

$$(6.2)\ P_n(k) = \lambda^k e^{-\lambda}/k!$$

Геометрическое распределение

Пусть производятся независимые испытания, в каждом из которых вероятность появления события A равна $p\ (0 < p < 1)$ и, следовательно, вероятность его непоявления $q = 1 - p$. Испытания заканчиваются, как только появится событие A. Таким образом, если событие A появилось в k-м испытании, то в предшествующих $k - 1$ испытаниях оно не появлялось.

Обозначим через X дискретную случайную величину – число испытаний, которые нужно провести до первого появления события A. Очевидно, возможными значениями X являются натуральные числа: $x_1 = 1, x_2 = 2, \ldots$

Пусть в первых $k - 1$ испытаниях событие A не наступило, а в k-м испытании появилось. Вероятность этого «сложного события», по теореме умножения вероятностей независимых событий,

$$(6.3)\ P(X = k) = q^{k-1}p$$

Полагая $k = 1, 2$, получим геометрическую прогрессию с первым членом p и знаменателем $q\ (0 < q < 1)$: $p, qp, q^2p, \ldots q^{k-1}p, \ldots$

По этой причине распределение (6.3) называют геометрическим.

Гипергеометрическое распределение

Пусть в партии из N изделий имеется M стандартных ($M < N$). Из партии случайно отбирают n изделий (каждое изделие может быть извлечено с одинаковой вероятностью), причем отобранное изделие перед отбором следующего не возвращается в партию (поэтому формула Бернулли здесь неприменима). Обозначим через X случайную величину – число m стандартных изделий среди n отобранных. Очевидно, возможные значения X таковы: $0, 1, 2, \ldots, min\ (M, n)$.

Найдем вероятность того, что $X = m$, т. е. что среди n отобранных изделий ровно m стандартных.

$$(6.4)\ P(X = m) = \frac{C_M^m C_{N-M}^{n-m}}{C_N^n}$$

Глава седьмая. МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Как уже известно, закон распределения полностью характеризует случайную величину. Однако часто закон распределения неизвестен и приходится ограничиваться меньшими сведениями. Иногда даже выгоднее пользоваться числами, которые описывают случайную величину суммарно; такие числа называют *числовыми характеристиками случайной величины*. К числу важных числовых характеристик относится математическое ожидание.

Математическим ожиданием дискретной случайной величины называют сумму произведений всех ее возможных значений на их вероятности.

$$(7.1)\ M(X) = \sum_{i=1}^{\infty} x_i p_i$$

Математическое ожидание приближенно равно (тем точнее, чем больше число испытаний) среднему арифметическому наблюдаемых значений случайной величины.

Математическое ожидание постоянной величины равно самой постоянной:

$$(7.2)\ M(C) = C$$

Постоянный множитель можно выносить за знак математического ожидания:

(7.3) $M(CX) = CM(X)$

Математическое ожидание произведения двух независимых случайных величин равно произведению их математических ожиданий:

(7.4) $M(XY) = M(X)M(Y)$

Математическое ожидание суммы двух случайных величин равно сумме математических ожиданий слагаемых:

(7.5) $M(X + Y) = M(X) + M(Y)$

Математическое ожидание $M(X)$ числа появлений события A в n независимых испытаниях равно произведению числа испытаний на вероятность появления события в каждом испытании:

(7.6) $M(X) = np$

Глава восьмая. ДИСПЕРСИЯ ДИСКРЕТНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Легко указать такие случайные величины, которые имеют одинаковые математические ожидания, но различные возможные значения. Зная лишь математическое ожидание случайной величины, еще нельзя судить ни о том, какие возможные значения она может принимать, ни о том, как они рассеяны вокруг математического ожидания. Другими словами, математическое ожидание полностью случайную величину не характеризует.

По этой причине наряду с математическим ожиданием вводят и другие числовые характеристики. Так, например, для того чтобы оценить, как рассеяны возможные значения случайной величины вокруг ее математического ожидания, пользуются, в частности, числовой характеристикой, которую называют дисперсией.

Отклонением называют разность между случайной величиной и ее математическим ожиданиям. Но, математическое ожидание отклонения равно нулю:

(8.1) $M[X - M(X)] = 0$

Дисперсией (рассеянием) дискретной случайной величины называют математическое ожидание квадрата отклонения случайной величины от ее математического ожидания:

(8.2) $D(X) = M[X - M(X)]^2$

Дисперсия равна разности между математическим ожиданием квадрата случайной величины X и квадратом ее математического ожидания:

(8.3) $D(X) = M(X^2) - [M(X)]^2$

Дисперсия постоянной величины C равна нулю:

(8.4) $D(C) = 0$

Постоянный множитель можно выносить за знак дисперсии, возводя его в квадрат:

(8.5) $D(CX) = C^2D(X)$

Дисперсия суммы двух независимых случайных величин равна сумме дисперсий этих величин:

(8.6) $D(X + Y) = D(X) + D(Y)$

Дисперсия разности двух независимых случайных величин равна сумме их дисперсий:

(8.7) $D(X - Y) = D(X) + D(Y)$

Дисперсия числа появлений события A в n независимых испытаниях, в каждом из которых вероятность p появления события постоянна, равна произведению числа испытаний на вероятности появления и непоявления события в одном испытании:

(8.8) $D(X) = npq$

Средним квадратическим отклонением случайной величины X называют квадратный корень из дисперсии:

(8.9) $\sigma(X) = \sqrt{D(X)}$

Математическое ожидание среднего арифметического одинаково распределенных взаимно независимых случайных величин равно математическому ожиданию a каждой из величин:

$$(8.10)\ M(\bar{X}) = a$$

Дисперсия среднего арифметического n одинаково распределенных взаимно независимых случайных величин в n раз меньше дисперсии D каждой из величин:

$$(8.11)\ D(\bar{X}) = D/n$$

Среднее квадратическое отклонение среднего арифметического n одинаково распределенных взаимно независимых случайных величин в $\sqrt{n}$ раз меньше среднего квадратического отклонения σ каждой из величин:

$$(8.12)\ \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Начальные и центральные теоретические моменты

Рассмотрим дискретную случайную величину X, заданную законом распределения:

X	1	2	5	100
P	0,6	0,2	0,19	0,01

Найдем математическое ожидание X: $M(X) = 1*0,6 + 2*0,2 + 5*0,19 +100*0,01 =2,95.$

Напишем закон распределения X^2:

X^2	1	4	25	10 000
p	0,6	0,2	0,19	0,01

Найдем математическое ожидание X^2: $M(X^2) = 1*0,6 + 4*0,2 + 25*0,19 + 10\ 000*0,01 = 106,15.$ Видим, что $M(X^2)$ значительно больше $M(X)$. Это объясняется тем, что после возведения в квадрат возможное значение величины X^2, соответствующее значению $x = 100$ величины X, стало равным 10 000, т. е. значительно увеличилось; вероятность же этого значения мала (0,01).

Таким образом, переход от $M(X)$ к $M(X^2)$ позволил лучше учесть влияние на математическое ожидание того возможного значения, которое велико и имеет малую вероятность. Разумеется, если бы величина X имела несколько больших и маловероятных значений, то переход к величине X^2, а тем более к величинам X^3, X^4 и т. д., позволил бы еще больше «усилить роль» этих больших, но маловероятных возможных значений. Вот почему оказывается целесообразным рассматривать математическое ожидание целой положительной степени случайной величины (не только дискретной, но и непрерывной).

Начальным моментом порядка k случайной величины X называют математическое ожидание величины X^k:

$$(8.13)\ \nu_k = M(X^k)$$

В частности,

$$(8.14)\ \nu_1 = M(X),\ \ \nu_2 = M(X^2)$$

Центральным моментом порядка k случайной величины X называют математическое ожидание величины $(X–M(X))^k$:

$$(8.15)\ \mu_k = M[(X - M(X))^k]$$

В частности,

$$(8.16)\ \mu_1 = M[X - M(X)] = 0, \qquad \mu_2 = M[(X - M(X))^2] = D(X)$$

Глава девятая. ЗАКОН БОЛЬШИХ ЧИСЕЛ

Неравенство Чебышева. Вероятность P того, что отклонение случайной величины X от ее математического ожидания по абсолютной величине меньше положительного числа ε, можно выразить неравенством:

$$(9.1)\ P(|X - M(X)| < \varepsilon) \geq \frac{1 - D(X)}{\varepsilon^2}$$

Теорема Чебышева. Если $X_1, X_2, \ldots, X_n, \ldots$ – попарно независимые случайные величины, причем дисперсии их равномерно ограничены (не превышают постоянного числа C), то, как бы мало ни было положительное число ε, вероятность неравенства

$$(9.2)\ \left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \frac{M(X_1) + M(X_2) + \cdots + M(X_n)}{n}\right| < \varepsilon$$

будет как угодно близка к единице, если число случайных величин достаточно велико.

Сущность теоремы Чебышева. Среднее арифметическое достаточно большого числа независимых случайных величин (дисперсии которых равномерно ограничены) утрачивает характер случайной величины. Объясняется это тем, что отклонения каждой из величин от своих математических ожиданий могут быть как положительными, так и отрицательными, а в среднем арифметическом они взаимно погашаются.

На теореме Чебышева основан широко применяемый в статистике выборочный метод, суть которого состоит в том, что по сравнительно небольшой случайной выборке судят обо всей совокупности (генеральной совокупности) исследуемых объектов.

Теорема Бернулли. Если в каждом из n независимых испытаний вероятность p появления события A постоянна, то как угодно близка к единице вероятность того, что отклонение относительной частоты от вероятности p по абсолютной величине будет сколь угодно малым, если число испытаний достаточно велико.

Другими словами, если ε – сколь угодно малое положительное число, то при соблюдении условий теоремы имеет место равенство

$$(9.3)\ \lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1$$

Глава десятая. ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Функцией распределения называют функцию $F(x)$, определяющую вероятность того, что случайная величина X в результате испытания примет значение, меньшее x, т. е.

$$(10.1)\ F(x) = P(X < x)$$

Геометрически это равенство можно истолковать так: $F(x)$ есть вероятность того, что случайная величина примет значение, которое изображается на числовой оси точкой, лежащей левее точки x. Иногда вместо термина «функция распределения» используют термин «интегральная функция».

Значения функции распределения принадлежат отрезку [0, 1]:

$$(10.2)\ 0 \le F(x) \le 1$$

$F(x)$ – неубывающая функция, т. е.

$$(10.3)\ F(x_2) \ge F(x_1), \text{если } x_2 > x_1$$

Вероятность того, что случайная величина примет значение, заключенное в интервале (a, b), равна приращению функции распределения на этом интервале:

$$(10.4)\ P(a \le X \le b) = F(b) - F(a)$$

Вероятность того, что непрерывная случайная величина X примет одно определенное значение, равна нулю.

Если возможные значения случайной величины принадлежат интервалу (a, b), то: 1) $F(x) = 0$ при $x \le a$ 2) $F(x) = 1$ при $x \ge b$.

Глава одиннадцатая. ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ НЕПРЕРЫВНОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Плотностью распределения вероятностей непрерывной случайной величины X называют функцию $f(x)$ – первую производную от функции распределения $F(x)$:

$$(11.1)\ f(x) = F'(x)$$

Вероятность того, что непрерывная случайная величина X примет значение, принадлежащее интервалу (a, b), равна определенному интегралу от плотности распределения, взятому в пределах от a до b:

$$(11.2)\ P(a < X < b) = \int_a^b f(x)dx$$

Зная плотность распределения $f(x)$, можно найти функцию распределения $F(x)$ по формуле

$$(11.3)\ F(x) = \int_{-\infty}^{x} f(x)dx$$

Глава двенадцатая. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Нормальным называют распределение вероятностей непрерывной случайной величины, которое описывается плотностью

$$(12.1)\ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Нормальное распределение определяется двумя параметрами a и σ. Достаточно знать эти параметры, чтобы задать нормальное распределение. Вероятностный смысл этих параметров таков: a есть математическое ожидание, σ – среднее квадратическое отклонение нормального распределения.

График плотности нормального распределения называют нормальной кривой (кривой Гаусса).

Исследуем функцию (12.1) методами дифференциального исчисления.

Функция определена на всей оси x. При всех значениях x функция принимает положительные значения, т. е. нормальная кривая расположена над осью Ox. Предел функции при неограниченном возрастании x (по абсолютной величине) равен нулю:

$$(12.2)\ \lim_{|x|\to\infty} y = 0$$

т.е. ось Ox служит горизонтальной асимптотой графика.

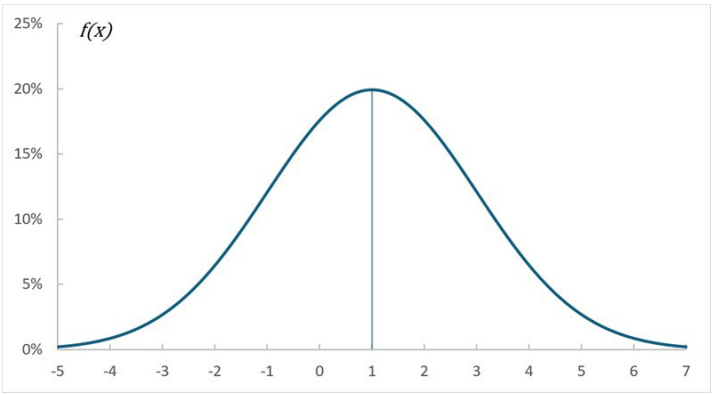

Рис. 3. Нормальная кривая при $a = 1$ и $\sigma = 2$

Исследуем функцию на экстремум. Найдем первую производную:

$$(12.3)\ y' = -\frac{x-a}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$y' = 0$ при $x = a$, $y' > 0$ при $x < a$, $y' < 0$ при $x > a$.

Следовательно, при x = a функция имеет максимум, равный:

$$(12.4)\ \frac{1}{\sigma\sqrt{2\pi}}$$

Разность $x - a$ содержится в аналитическом выражении функции в квадрате, т. е. график функции симметричен относительно прямой $x - a$.

Исследуем функцию на точки перегиба. Найдем вторую производную:

$$(12.5)\ y'' = -\frac{1}{\sigma^3\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} \left[1 - \frac{(x-a)^2}{\sigma^2}\right]$$

Легко видеть, что при $x = a + \sigma$ и $x = a - \sigma$ вторая производная равна нулю, а при переходе через эти точки она меняет знак. В обеих этих точках значение функции равно:

$$(12.6)\ \frac{1}{\sigma\sqrt{2\pi e}}$$

Таким образом, точки графика

$$(12.7)\ \left(a - \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right) \text{ и } \left(a + \sigma, \frac{1}{\sigma\sqrt{2\pi e}}\right)$$

являются точками перегиба.

Формулировка центральной предельной теоремы

Известно, что нормально распределенные случайные величины широко распространены на практике. Чем это объясняется? Ответ на этот вопрос был дан выдающимся русским математиком А. М. Ляпуновым (центральная предельная теорема): если случайная величина X представляет собой сумму очень большого числа взаимно независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то X имеет распределение, близкое к нормальному.

Характеристической функцией случайной величины X называют функцию

$$(12.8)\ \varphi(t) = M[e^{itX}]$$

Для дискретной случайной величины X с возможными значениями x_k и их вероятностями p_k характеристическая функция

$$(12.9)\ \varphi(t) = \sum_k e^{itx} p_k$$

Для непрерывной случайной величины X с плотностью распределения $f(x)$ характеристическая функция

$$(12.10)\ \varphi(t) = \int_{-\infty}^{\infty} e^{itx} f(x) dx$$

Оценка отклонения теоретического распределения от нормального. Асимметрия и эксцесс

Эмпирическим называют распределение относительных частот. Эмпирические распределения изучает математическая статистика.

Теоретическим называют распределение вероятностей. Теоретические распределения изучает теория вероятностей. В этом параграфе рассматриваются теоретические распределения.

При изучении распределений, отличных от нормального, возникает необходимость количественно оценить это различие. С этой целью вводят специальные характеристики, в частности асимметрию и эксцесс. Для нормального распределения эти характеристики равны нулю. Поэтому если для изучаемого распределения асимметрия и эксцесс имеют небольшие значения, то можно предположить близость этого распределения к нормальному. Наоборот, большие значения асимметрии и эксцесса указывают на значительное отклонение от нормального.

Асимметрией теоретического распределения называют отношение центрального момента третьего порядка к кубу среднего квадратического отклонения:

$$(12.11)\ A_s = \frac{\mu_3}{\sigma^3}$$

Асимметрия положительна, если «длинная часть» кривой распределения расположена справа от математического ожидания; асимметрия отрицательна, если «длинная часть» кривой расположена слева от математического ожидания.

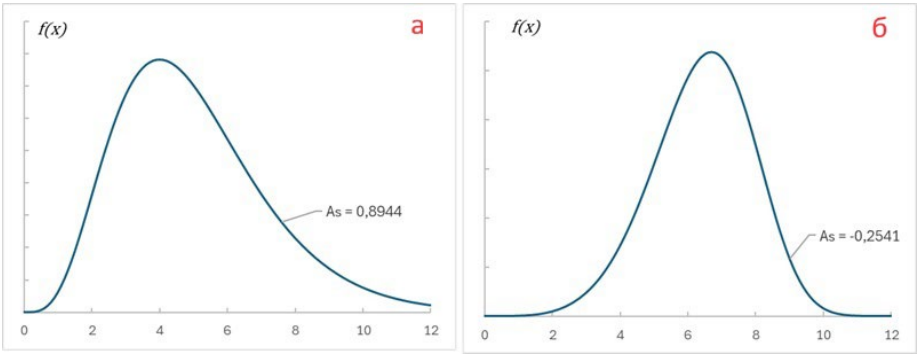

Рис. 4. Распределения а) с правым хвостом на основе гамма распределения, б) с левым хвостом на основе распределения Вейбулла (формулы см. в приложенном Excel-файле)

Для оценки «крутости», т. е. большего или меньшего подъема кривой теоретического распределения по сравнению с нормальной кривой, пользуются характеристикой — эксцессом.

Эксцессом теоретического распределения называют характеристику, которая определяется равенством

$$(12.12)\ E_k = \frac{\mu_4}{\sigma^4} - 3$$

Для нормального распределения эксцесс равен нулю. Если эксцесс положительный, то кривая имеет более высокую и «острую» вершину, чем нормальная (рис. 5а); если эксцесс отрицательный, то сравниваемая кривая имеет более низкую и «плоскую» вершину, чем нормальная (рис. 5б).

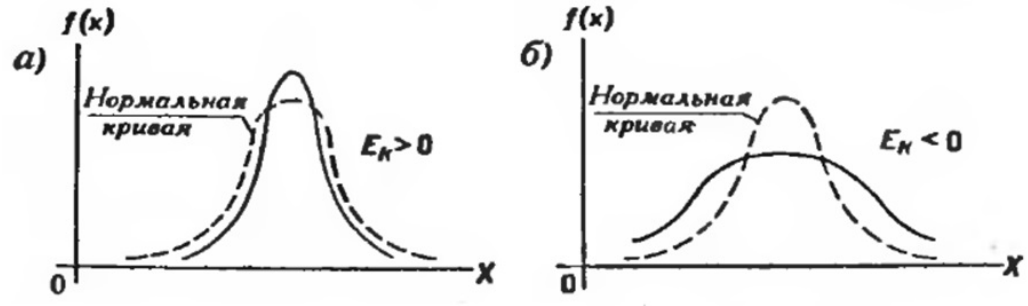

Рис. 5. Эксцесс кривых распределения: а) больше нуля, б) меньше нуля

На самом деле это не так. Например, положительный эксцесс равный трем имеют, как t-распределение Стьюдента с $df = 6$, так и стандартное распределение Лапласа:

$$(12.13)\ f(x) = \frac{1}{2e^{|x|}}$$

При том, что первое имеет более плоскую вершину, а второе – более острую, в сравнении с нормальным распределением (см. также Эксцесс распределения случайной величины):

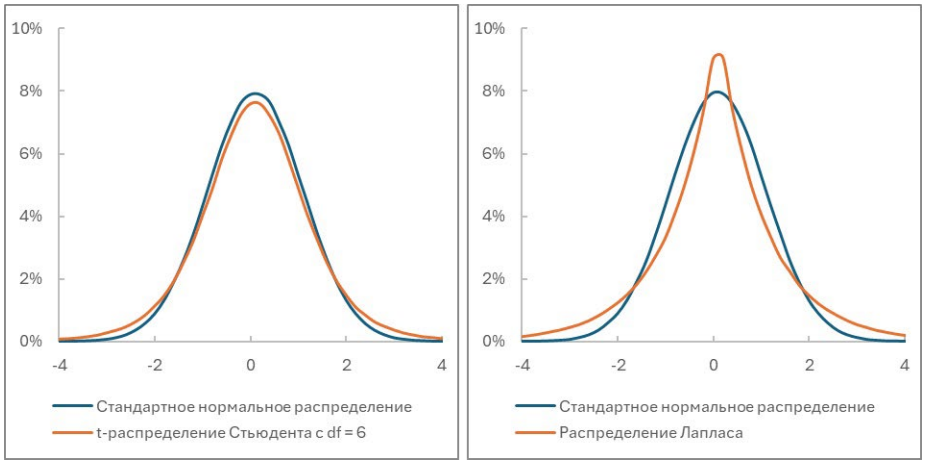

Рис. 6. Влияние толстых хвостов: а) t-распределение Стьюдента с $df = 6$; б) стандартное распределение Лапласа; в обоих случаях $E_k = 3$

Поскольку эксцесс пропорционален четвертой степени отклонения случайной величины от среднего, то возрастает роль больших отклонений, т.е. хвоста. Положительный эксцесс обусловлен более тяжелыми хвостами относительно нормального распределения, отрицательный – более легкими.

Функция одного случайного аргумента и ее распределение

Если каждому возможному значению случайной величины X соответствует одно возможное значение случайной величины Y, то Y называют функцией случайного аргумента X: $Y = \varphi(X)$. Как найти распределение функции $Y = \varphi(X)$, зная плотность распределения случайного аргумента X? Доказано: если $y = \varphi(x)$ – дифференцируемая строго возрастающая или строго убывающая функция, обратная функция которой $x = \psi(y)$, то плотность распределения $g(y)$ случайной величины Y

$$(12.14)\ g(y) = f[\psi(y)]|\psi'(y)$$

Распределение суммы независимых слагаемых. Устойчивость нормального распределения

Плотность распределения суммы независимых случайных величин называют *композицией*. Закон распределения вероятностей называют *устойчивым*, если композиция таких законов есть тот же закон (отличающийся, вообще говоря, параметрами). Нормальный закон обладает свойством устойчивости: композиция нормальных законов также имеет нормальное распределение (математическое ожидание и дисперсия этой композиции равны соответственно суммам математических ожиданий и дисперсий слагаемых).

Например, если X и Y — независимые случайные величины, распределенные нормально с математическими ожиданиями и дисперсиями, соответственно равными $a_1 = 3$, $a_2 = 4$, $D_1 = 1$, $D_2 = 0{,}5$, то композиция этих величин (т. е. плотность вероятности суммы $Z = X + Y$) также распределена нормально, причем математическое ожидание и дисперсия композиции соответственно равны $a = 3 + 4 = 7$; $D = 1 + 0{,}5 = 1{,}5$.

Распределение «хи квадрат»

Пусть X_i $(i = 1, 2, \ldots, n)$ – нормальные независимые случайные величины, причем математическое ожидание каждой из них равно нулю, а среднее квадратическое отклонение – единице. Тогда сумма квадратов этих величин

$$(12.15)\ \chi^2 = \sum_{i=1}^{n} X_i^2$$

распределена по закону χ^2 («хи квадрат») с $k = n$ степенями свободы; если же эти величины связаны одним линейным соотношением, например $\Sigma X_i = n\bar{X}$, то число степеней свободы $k = n - 1$.

Плотность этого распределения

$$(12.16)\ f(x) = \begin{cases} 0 & \text{при } x \le 0 \\ \dfrac{1}{2^{k/2}\Gamma(k/2)} e^{-x/2} x^{(k/2)-1} & \text{при } x > 0 \end{cases}$$

где…

$$(12.17)\ \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

… есть гамма-функция; в частности, $\Gamma(n+1) = n!$

Отсюда видно, что распределение «хи квадрат» определяется одним параметром – числом степеней свободы k. С увеличением числа степеней свободы распределение медленно приближается к нормальному.

Распределение Стьюдента

Пусть Z – нормальная случайная величина, причем $M(Z) = 0$, $\sigma(Z) = 1$, а V – независимая от Z величина, которая распределена по закону χ^2 с k степенями свободы. Тогда величина

$$(12.18)\ \mathrm{T} = \frac{Z}{\sqrt{V/k}}$$

имеет распределение, которое называют t-распределением или распределением Стьюдента (псевдоним английского статистика В. Госсета), с k степенями свободы.

Итак, отношение нормированной нормальной величины к квадратному корню из независимой случайной величины, распределенной по закону «хи квадрат» с k степенями свободы, деленной на k, распределено по закону Стьюдента с k степенями свободы. С возрастанием числа степеней свободы распределение Стьюдента быстро приближается к нормальному.

Распределение F Фишера — Снедекора

Если U и V – независимые случайные величины, распределенные по закону χ^2 со степенями свободы k_1 и k_2, то величина

$$(12.19)\ F = \frac{U/k_1}{U/k_2}$$

имеет распределение, которое называют распределением F Фишера–Снедекора со степенями свободы k_1 и k_2 (иногда его обозначают через V^2).

Плотность этого распределения

$$(12.20)\ f(x) = \begin{cases} 0 & \text{при } x \le 0 \\ C_0 \dfrac{x^{(k_1-2)/2}}{(k_2+k_1x)^{(k_1+k_2)/2}} & \text{при } x > 0 \end{cases}$$

где

$$(12.21)\ C_0 = \frac{\Gamma\left(\frac{k_1+k_2}{2}\right) k_1^{k_1/2} k_2^{k_2/2}}{\Gamma(k_1/2)\Gamma(k_2/2)}$$

Мы видим, что распределение F определяется двумя параметрами – числами степеней свободы.

Глава тринадцатая. ПОКАЗАТЕЛЬНОЕ РАСПРЕДЕЛЕНИЕ

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

$$(13.1)\ f(x) = \begin{cases} 0 & \text{при } x < 0 \\ \lambda e^{-\lambda x} & \text{при } x \ge 0 \end{cases}$$

где λ – постоянная положительная величина.

Показательное распределение определяется одним параметром λ. Эта особенность показательного распределения указывает на его преимущество по сравнению с распределениями, зависящими от большего числа параметров. Обычно параметры неизвестны и приходится находить их оценки (приближенные значения); разумеется, проще оценить один параметр, чем два или более.

Функция распределения показательного закона:

$$(13.2)\ F(x) = \int_{-\infty}^{x} f(x)dx = \int_{-\infty}^{0} 0dx + \lambda \int_{0}^{x} e^{-\lambda x}dx = 1 - e^{-\lambda x}$$

Математическое ожидание показательного распределения $M(X) = 1/\lambda$. Среднее квадратическое отклонение $\sigma(X) = 1/\lambda$.

Глава четырнадцатая. СИСТЕМА ДВУХ СЛУЧАЙНЫХ ВЕЛИЧИН

До сих пор рассматривались случайные величины, возможные значения которых определялись одним числом. Такие величины называют одномерными. Кроме одномерных случайных величин изучают величины, возможные значения которых определяются двумя, тремя, …, n числами.

Будем обозначать через (X, Y) двумерную случайную величину. Каждую из величин X и Y называют составляющей (компонентой); обе величины X и Y, рассматриваемые одновременно, образуют систему двух случайных величин.

Рассмотрим двумерную случайную величину (X, Y). Пусть x, y – пара действительных чисел. Вероятность события, состоящего в том, что X примет значение, меньшее x, и при этом Y примет значение, меньшее y, обозначим через $F(x,y)$. Если x и y будут изменяться, то, вообще говоря, будет изменяться и $F(x,y)$, т. е. $F(x,y)$ есть функция от x и y.

Функцией распределения двумерной случайной величины (X, Y) называют функцию $F(x,y)$, определяющую для каждой пары чисел x, y вероятность того, что X примет значение, меньшее x, и при этом Y примет значение, меньшее y: $F(x,y) = P(X<x, Y<y)$. Геометрически это равенство можно истолковать так: $F(x,y)$ есть вероятность того, что случайная точка (X, Y) попадет в бесконечный квадрант с вершиной (x, y), расположенный левее и ниже этой вершины.

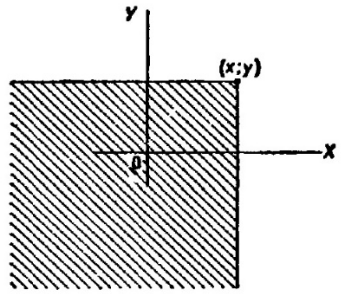

Рис. 7. Геометрическое толкование функции распределения двумерной случайной величины

Числовые характеристики системы двух случайных величин. Корреляционный момент. Коэффициент корреляции

Для описания системы двух случайных величин кроме математических ожиданий и дисперсий составляющих используют и другие характеристики; к их числу относятся корреляционный момент и коэффициент корреляции.

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

Корреляционный момент служит для характеристики связи между величинами X и Y. Корреляционный момент равен нулю, если X и Y независимы; следовательно, если корреляционный момент не равен нулю, то X и Y – зависимые случайные величины.

Коэффициентом корреляции r_{xy} случайных величин X и Y называют отношение корреляционного момента к произведению средних квадратических отклонений этих величин:

$$(14.1)\ r_{xy} = \mu_{xy}/(\sigma_x\sigma_y)$$

Так как размерность μ_{xy}, равна произведению размерностей величин X и Y, σ_x имеет размерность величины X, σ_y имеет размерность величины Y, то r_{xy} – безразмерная величина. Таким образом, величина коэффициента корреляции не зависит от выбора единиц измерения случайных величин. В этом состоит преимущество коэффициента корреляции перед корреляционным моментом.

Линейная регрессия. Прямые линии среднеквадратической регрессии

Рассмотрим двумерную случайную величину (X, Y), где X и Y – зависимые случайные величины. Представим одну из величин как функцию другой. Ограничимся приближенным представлением (точное приближение, вообще говоря, невозможно) величины Y в виде линейной функции величины X: $Y \simeq g(x) = \alpha X + \beta$, где α и β – параметры, подлежащие определению. Это можно сделать различными способами: наиболее употребительный из них – метод наименьших квадратов.

Функцию $g(x) = \alpha X + \beta$ называют «наилучшим приближением» Y в смысле метода наименьших квадратов, если математическое ожидание $M[Y - g(X)]^2$ принимает наименьшее возможное значение. Функцию $g(x)$ называют *среднеквадратической регрессией* Y на X, которая имеет вид

$$(14.2)\ g(X) = m_y + r\frac{\sigma_y}{\sigma_x}(X - m_x)$$

где

$$(14.3)\ m_x = M(X),\ m_y = M(Y),\ \sigma_x = \sqrt{D(X)}, \sigma_y = \sqrt{D(Y)}, r = \frac{\mu_{xy}}{\sigma_x \sigma_y}$$

Коэффициент $\beta = r\sigma_y/\sigma_x$ — называют коэффициентом регрессии Y на X, а прямую

$$(14.4)\ y - m_y = r\frac{\sigma_y}{\sigma_x}(X - m_x)$$

называют прямой среднеквадратической регрессии Y на X.

Величину

$$(14.5)\ \sigma_y^2(1 - r^2)$$

называют остаточной дисперсией случайной величины Y относительно случайной величины X; она характеризует величину ошибки, которую допускают при замене Y линейной функцией $g(X)$. При $r = \pm 1$ остаточная дисперсия равна нулю; другими словами, при этих крайних значениях коэффициента корреляции не возникает ошибки при представлении Y в виде линейной функции от X.

ЧАСТЬ ТРЕТЬЯ. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Глава пятнадцатая. ВЫБОРОЧНЫЙ МЕТОД

Первая задача математической статистики – указать способы сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов. Вторая задача математической статистики – разработать методы анализа статистических данных в зависимости от целей исследования. Сюда относятся:

a) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости случайной величины от одной или нескольких случайных величин и др.;
b) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Современную математическую статистику определяют как *науку о принятии решений в условиях неопределенности*.

Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 – n_2 раз, x_k – n_k раз и $\Sigma n_i = n$ – объем выборки. Наблюдаемые значения x_i называют вариантами, а последовательность вариант, записанных в возрастающем порядке, – вариационным рядом. Числа наблюдений называют частотами, а их отношения к объему выборки $n_i/n = W_i$ – *относительными частотами*.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения x относительную частоту события $X < x$. $F^*(x) = n_x/n$.

В отличие от эмпирической функции распределения выборки функцию распределения $F(x)$ генеральной совокупности называют *теоретической функцией распределения*. Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция $F(x)$ определяет вероятность события $X < x$, а эмпирическая функция $F^*(x)$ определяет относительную частоту этого же события. При больших n числа $F^*(x)$ и $F(x)$ мало отличаются одно от другого. Отсюда следует целесообразность использования эмпирической функции распределения выборки для приближенного представления теоретической (интегральной) функции распределения генеральной совокупности.

Глава шестнадцатая. СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Статистической оценкой неизвестного параметра теоретического распределения называют функцию от наблюдаемых случайных величин.

Для того чтобы статистические оценки давали «хорошие» приближения оцениваемых параметров, они должны удовлетворять определенным требованиям.

Пусть Θ^* – статистическая оценка неизвестного параметра Θ теоретического распределения. Допустим, что по выборке объема n найдена оценка Θ_1^*. Повторим опыт, т. е. извлечем из генеральной совокупности другую выборку того же объема и по ее данным найдем оценку Θ_2^*. Повторяя опыт многократно, получим числа Θ_1^*, Θ_2^*, ..., Θ_k^*, которые, вообще говоря, различны между собой. Таким образом, оценку Θ^* можно рассматривать как случайную величину, а числа Θ_1^*, Θ_2^*, ..., Θ_k^* – как ее возможные значения.

Представим себе, что оценка Θ^* дает приближенное значение Θ. *Несмещенной* называют статистическую оценку Θ^*, математическое ожидание которой равно оцениваемому параметру Θ при любом объеме выборки, т. е. $M[\Theta^*] = \Theta$.

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако было бы ошибочным считать, что несмещенная оценка всегда дает хорошее приближение оцениваемого параметра. Действительно, возможные значения Θ^* могут быть сильно рассеяны вокруг своего среднего значения, т. е. дисперсия $D(\Theta^*)$ может быть значительной. В этом случае найденная по данным одной выборки оценка, например Θ_1^*, может оказаться весьма удаленной от среднего значения $\bar{\Theta}_1^*$, а значит, и от самого оцениваемого параметра Θ; приняв Θ_1^* в качестве приближенного значения Θ, мы допустили бы большую ошибку.

Эффективной называют статистическую оценку, которая (при заданном объеме выборки n) имеет наименьшую возможную дисперсию.

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

Если по нескольким выборкам достаточно большого объема из одной и той же генеральной совокупности будут найдены выборочные средние, то они будут приближенно равны между собой. В этом состоит свойство *устойчивости выборочных средних*.

Групповая, внутригрупповая, межгрупповая и общая дисперсии

Допустим, что все значения количественного признака X совокупности, безразлично — генеральной или выборочной, разбиты на k групп. Рассматривая каждую группу как самостоятельную совокупность, можно найти групповую среднюю и дисперсию значений признака, принадлежащих группе, относительно групповой средней.

Групповой дисперсией называют дисперсию значений признака, принадлежащих группе, относительно групповой средней

$$(16.1)\ D_{\text{гр}} = \frac{\sum n_i (x_i - \bar{x}_j)^2}{N_j}$$

где n_i – частота значения x_i, j – номер группы; $\bar{x}_j$ – групповая средняя группы j; $N_j = \Sigma n_i$ – объем группы j.

Зная дисперсию каждой группы, можно найти их среднюю арифметическую.

Внутригрупповой дисперсией называют среднюю арифметическую дисперсий, взвешенную по объемам групп:

$$(16.2)\ D_{\text{внгр}} = \frac{\sum N_j D_{j\text{гр}}}{n}$$

где N_j – объем группы j, $n = \Sigma N_j$ – объем всей совокупности.

Зная групповые средние и общую среднюю, можно найти дисперсию групповых средних относительно общей средней.

Межгрупповой дисперсией называют дисперсию групповых средних относительно общей средней:

$$(16.3)\ D_{\text{межгр}} = \frac{\sum N_j (\bar{x}_j - \bar{x})^2}{n}$$

где $\bar{x}_j$ – групповая средняя группы j; N_j – объем группы j; $\bar{x}$ – общая средняя; n – объем всей совокупности.

Общей дисперсией называют дисперсию значений признака всей совокупности относительно общей средней:

$$(16.4)\ D_{\text{общ}} = \frac{\sum n_i (x_i - \bar{x})^2}{n}$$

где n_i – частота значения x_i, $\bar{x}$ – общая средняя; n – объем всей совокупности.

Если совокупность состоит из нескольких групп, то общая дисперсия равна сумме внутригрупповой и межгрупповой дисперсий: $D_{общ} = D_{внгр} + D_{межгр}$.

Оценка генеральной дисперсии по исправленной выборочной

Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка будет приводить к систематическим ошибкам, давая заниженное значение генеральной дисперсии. Выборочная дисперсия является смещенной оценкой $D_г$. В качестве оценки генеральной дисперсии принимают исправленную дисперсию

$$(16.5)\ s^2 = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_{\text{в}})^2}{n-1}$$

Подчеркнем, что s не является несмещенной оценкой; чтобы отразить этот факт, мы написали «исправленное» среднее квадратическое отклонение.

Точность оценки, доверительная вероятность (надежность). Доверительный интервал

Точечной называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, – точечные. При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

Пусть найденная по данным выборки статистическая характеристика Θ^* служит оценкой неизвестного параметра Θ. Будем считать Θ постоянным числом (Θ может быть и случайной величиной). Ясно, что Θ^* тем точнее определяет параметр Θ, чем меньше абсолютная величина разности $|\Theta - \Theta^*|$. Другими словами, если $\delta > 0$ и $|\Theta - \Theta^*| < \delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует *точность оценки*.

Однако статистические методы не позволяют категорически утверждать, что оценка Θ^* удовлетворяет неравенству $|\Theta - \Theta^*| < \delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки Θ по Θ^* называют вероятность γ, с которой осуществляется неравенство $|\Theta - \Theta^*| < \delta$. Обычно надежность оценки задается наперед, причем в

качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что $|\Theta - \Theta^*| < \delta$, равна γ

$$(16.6)\ P[|\Theta - \Theta^*| < \delta] = \gamma$$

Это соотношение следует понимать так: вероятность того, что интервал ($\Theta^* - \delta$, $\Theta^* + \delta$) заключает в себе (покрывает) неизвестный параметр Θ, равна γ.

Доверительным называют интервал ($\Theta^* - \delta$, $\Theta^* + \delta$), который покрывает неизвестный параметр с заданной надежностью γ.

Так как случайной величиной является не оцениваемый параметр Θ, а доверительный интервал, то более правильно говорить не о вероятности попадания Θ в доверительный интервал, а о вероятности того, что доверительный интервал покроет Θ.

Метод доверительных интервалов разработал американский статистик Ю. Нейман, исходя из идей английского статистика Р. Фишера.

Доверительные интервалы для оценки математического ожидания нормального распределения при известном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}$.

$$(16.7)\ P\left(\frac{\bar{x} - t\sigma}{\sqrt{n}} < a < \frac{\bar{x} + t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma$$

где γ – заданная надежность, $t = \delta\sqrt{n}/\sigma$.

Смысл полученного соотношения таков: с надежностью γ можно утверждать, что доверительный интервал ($\bar{x} - t\sigma/\sqrt{n}$, $\bar{x} + t\sigma/\sqrt{n}$) покрывает неизвестный параметр а; точность оценки $\delta = t\sigma/\sqrt{n}$.

При возрастании объема выборки n число δ убывает и, следовательно, точность оценки увеличивается. Увеличение надежности оценки $\gamma = 2\Phi(t)$ приводит к увеличению t ($\Phi(t)$ – возрастающая функция), следовательно, и к возрастанию δ; другими словами, увеличение надежности влечет за собой уменьшение ее точности.

Пример. Случайная величина X имеет нормальное распределение с известным средним квадратическим отклонением $\sigma = 3$. Найти доверительные интервалы для оценки неизвестного математического ожидания a по выборочным средним $\bar{x}$, если объем выборки $n = 36$ и задана надежность оценки $\gamma = 0{,}95$.

Решение. Найдем t. Из соотношения $2\Phi(t) = 0{,}95$ получим $\Phi(t) = 0{,}475$. По таблицам находим $t = 1{,}96$.

Найдем точность оценки:

$$(16.8)\ \delta = \frac{t\sigma}{\sqrt{n}} = \frac{1{,}96 \cdot 3}{\sqrt{36}} = 0{,}98$$

Доверительный интервал таков: ($\overline{x}$ – 0,98; $\overline{x}$ + 0,98). Например, если $\overline{x}$ = 4,1, то доверительный интервал имеет следующие доверительные границы: $\overline{x}$ – 0,98 = 4,1 – 0,98 = 3,12; $\overline{x}$ + 0,98 = 4,1 + 0,98 = 5,08.

Таким образом, значения неизвестного параметра a, согласующиеся с данными выборки, удовлетворяют неравенству $3{,}12 < a < 5{,}08$. Подчеркнем, что было бы ошибочным написать $P(3{,}12 < a < 5{,}08) = 0{,}95$. Действительно, так как a – постоянная величина, то либо она заключена в найденном интервале (тогда событие $3{,}12 < a < 5{,}08$ достоверно и его вероятность равна единице), либо в нем не заключена (в этом случае событие $3{,}12 < a < 5{,}08$ невозможно и его вероятность равна нулю). Другими словами, доверительную вероятность не следует связывать с оцениваемым параметром; она связана лишь с границами доверительного интервала, которые, как уже было указано, изменяются от выборки к выборке.

Поясним смысл, который имеет заданная надежность. Надежность γ = 0,95 указывает, что если произведено достаточно большое число выборок, то 95% из них определяет такие доверительные

интервалы, в которых параметр действительно заключен; лишь в 5% случаев он может выйти за границы доверительного интервала.

Доверительные интервалы для оценки математического ожидания нормального распределения при неизвестном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительных интервалов. Разумеется, невозможно воспользоваться результатами предыдущего параграфа, в котором σ предполагалось известным.

Оказывается, что по данным выборки можно построить случайную величину (ее возможные значения будем обозначать через t):

$$(16.9)\ T = \frac{\bar{X} - a}{S/\sqrt{n}}$$

которая имеет распределение Стьюдента с $k = n - 1$ степенями свободы; здесь $\bar{X}$ – выборочная средняя, S – «исправленное» среднее квадратическое отклонение, n – объем выборки.

Плотность распределения Стьюдента

$$(16.10)\ S(t,n) = B_n\left[1 + \frac{t^2}{n-1}\right]^{-\frac{n}{2}}, \text{где } B_n = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\cdot\Gamma\left(\frac{n-1}{2}\right)}$$

Мы видим, что распределение Стьюдента определяется параметром n – объемом выборки (или числом степеней свободы $k = n - 1$) и не зависит от неизвестных параметров a и σ; эта особенность является его большим достоинством. Поскольку $S(t,n)$ – четная функция от t, вероятность осуществления неравенства

$$(16.11)\ \left|\frac{\bar{X} - a}{S/\sqrt{n}}\right| < \gamma$$

определяется

$$(16.12)\ P\left(\left|\frac{\bar{X} - a}{S/\sqrt{n}}\right| < t_\gamma\right) = 2\int_0^{t_\gamma} S(t,n)dt = \gamma$$

Заменив неравенство в круглых скобках равносильным ему двойным неравенством, получим

$$(16.13)\ P\left(\bar{X} - t_\gamma S/\sqrt{n} < a < \bar{X} + t_\gamma S/\sqrt{n}\right) = \gamma$$

Итак, пользуясь распределением Стьюдента, мы нашли доверительный интервал, покрывающий. неизвестный параметр a с надежностью γ. Здесь случайные величины $\bar{X}$ и S заменены неслучайными величинами $\bar{x}$ и s, найденными по выборке. По таблице по заданным n и γ можно найти t_γ.

n	γ			n	γ		
	0,95	0,99	0,999		0,95	0,99	0,999
5	2,7764	4,6041	8,6103	20	2,0930	2,8609	3,8834
6	2,5706	4,0321	6,8688	25	2,0639	2,7969	3,7454
7	2,4469	3,7074	5,9588	30	2,0452	2,7564	3,6594
8	2,3646	3,4995	5,4079	35	2,0322	2,7284	3,6007
9	2,3060	3,3554	5,0413	40	2,0227	2,7079	3,5581
10	2,2622	3,2498	4,7809	45	2,0154	2,6923	3,5258
11	2,2281	3,1693	4,5869	50	2,0096	2,6800	3,5004
12	2,2010	3,1058	4,4370	60	2,0010	2,6618	3,4632
13	2,1788	3,0545	4,3178	70	1,9949	2,6490	3,4372
14	2,1604	3,0123	4,2208	80	1,9905	2,6395	3,4180
15	2,1448	2,9768	4,1405	90	1,9870	2,6322	3,4032
16	2,1314	2,9467	4,0728	100	1,9842	2,6264	3,3915
17	2,1199	2,9208	4,0150	110	1,9820	2,6217	3,3820
18	2,1098	2,8982	3,9651	120	1,9801	2,6178	3,3742
19	2,1009	2,8784	3,9216	∞	1,9600	2,5758	3,2905

Рис. 8. Таблица значений $t_\gamma = t(\gamma, n)$

В Excel значение можно найти с помощью формулы =СТЬЮДЕНТ.ОБР(1-(1-γ)/2;n-1).

При неограниченном возрастании объема выборки n распределение Стьюдента стремится к нормальному. Значения в таблице для $n = \infty$ найдены по формуле =НОРМ.СТ.ОБР(1-(1- γ)/2). Поэтому практически при $n > 30$ можно вместо распределения Стьюдента пользоваться нормальным распределением.

Для малых выборок ($n < 30$) замена распределения нормальным приводит к грубым ошибкам, а именно к неоправданному сужению доверительного интервала, т. е. к повышению точности оценки. Например, если $n = 5$ и $\gamma = 0{,}99$, то, пользуясь распределением Стьюдента, найдем $t_\gamma = 4{,}6$, а используя функцию Лапласа, найдем $t_\gamma = 2{,}58$, т. е. доверительный интервал в последнем случае окажется более узким, чем найденный по распределению Стьюдента.

То обстоятельство, что распределение Стьюдента при малой выборке дает не вполне определенные результаты (широкий доверительный интервал), вовсе не свидетельствует о слабости метода Стьюдента, а объясняется тем, что малая выборка, разумеется, содержит малую информацию об интересующем нас признаке.

Метод моментов для точечной оценки параметров распределения

Можно доказать, что начальные и центральные эмпирические моменты являются состоятельными оценками соответственно начальных и центральных теоретических моментов того же порядка. На этом основан метод моментов, предложенный К. Пирсоном. Достоинство метода — сравнительная его простота. Метод моментов точечной оценки неизвестных параметров заданного распределения состоит в приравнивании теоретических моментов рассматриваемого распределения соответствующим эмпирическим моментам того же порядка.

А. Оценка одного параметра. Пусть задан вид плотности распределения $f(x, \Theta)$, определяемой одним неизвестным параметром Θ. Требуется найти точечную оценку параметра Θ.

Для оценки одного параметра достаточно иметь одно уравнение относительно этого параметра. Следуя методу моментов, приравняем, например, начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка: $\nu_1 = M_1$. Учитывая, что $\nu_1 = M(X)$, $M_1 = \bar{x}_B$, получим

$$(16.14)\ M(X) = \bar{x}_B$$

Математическое ожидание $M(X)$, как видно из соотношения

$$(16.15)\ M(X) = \int_{-\infty}^{\infty} xf(x;\Theta)dx = \varphi(x),$$

есть функция от Θ, поэтому (16.13) можно рассматривать как уравнение с одним неизвестным Θ. Решив это уравнение относительно параметра Θ, тем самым найдем его точечную оценку Θ^*, которая является функцией от выборочной средней, следовательно, и от вариант выборки:

$$(16.16)\ \Theta^* = \psi(x_1, x_2, \dots, x_n)$$

*Пример***.** Найти методом моментов по выборке $x_1, x_2, \dots, x_n$ точечную оценку неизвестного параметра λ показательного распределения, плотность распределения которого $f(x) = \lambda e^{-\lambda x}$ $(x \geq 0)$.

Решение. Приравняем начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка: $\nu_1 = M_1$. Учитывая, что $\nu_1 = M(X)$, $M_1 = \bar{x}_B$, получим $M(X) = \bar{x}_B$. Приняв во внимание, что математическое ожидание показательного распределения равно $1/\lambda$, имеем $1/\lambda = \bar{x}_B$. Отсюда $\lambda = 1/\bar{x}_B$. Итак, искомая точечная оценка параметра λ, показательного распределения равна величине, обратной выборочной средней: $\lambda^* = 1/\bar{x}_B$.

Б. Оценка двух параметров. Пусть задан вид плотности распределения $f(x, \Theta_1, \Theta_2)$, определяемой неизвестными параметрами Θ_1 и Θ_2. Для отыскания двух параметров необходимы два уравнения относительно этих параметров. Следуя методу моментов, приравняем, например, начальный теоретический момент первого порядка начальному эмпирическому моменту первого порядка и центральный теоретический момент второго порядка центральному эмпирическому моменту второго порядка: $\nu_1 = M_1$, $\mu_2 = M_2$.

Учитывая, что $\nu_1 = M(X)$, $\mu_2 = D(X)$, $M_1 = \bar{x}_B$, $m_2 = D_B$, получим $M(X) = \bar{x}_B$, $D(X) = D_B$. Математическое ожидание и дисперсия есть функции от Θ_1 и Θ_2, поэтому два последних равенства можно рассматривать как систему двух уравнений с двумя неизвестными Θ_1 и Θ_2. Решив эту систему относительно неизвестных параметров, тем самым получим их точечные оценки Θ^*_1 и Θ^*_2. Эти оценки являются функциями от вариант выборки:

$$(16.17)\ \Theta_1^* = \psi_1(x_1, x_2, \dots, x_n),\ \Theta_2^* = \psi_2(x_1, x_2, \dots, x_n)$$

Пример. Найти методом моментов по выборке $x_1, x_2, \dots, x_n$ точечные оценки неизвестных параметров a и σ нормального распределения

$$(16.18)\ f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{(2\sigma)^2}}$$

Решение. Приравняем начальные теоретические и эмпирические моменты первого порядка, а также центральные и эмпирические моменты второго порядка: $\nu_1 = M_1$, $\mu_2 = M_2$. Учитывая, что $\nu_1 = M(X)$, $\mu_2 = D(X)$, $M_1 = \bar{x}_B$, $m_2 = D_B$, получим $M(X) = \bar{x}_B$, $D(X) = D_B$. Приняв во внимание, что математическое ожидание нормального распределения равно параметру a, дисперсия равна σ^2, имеем: $a = \bar{x}_B$, $\sigma^2 = D_B$. Итак, искомые точечные оценки параметров нормального распределения: $a^* = \bar{x}_B$, $\sigma^* = \sqrt{D_B}$.

Метод наибольшего правдоподобия

Кроме метода моментов существуют и другие методы точечной оценки неизвестных параметров распределения. К ним относится метод *наибольшего правдоподобия*, предложенный Р. Фишером.

А. Дискретные случайные величины. Пусть X – дискретная случайная величина, которая в результате n испытаний приняла значения $x_1, x_2, \dots, x_n$. Допустим, что вид закона распределения величины X задан, но неизвестен параметр Θ, которым определяется этот закон. Требуется найти его точечную оценку.

Обозначим вероятность того, что в результате испытания величина X примет значение x_i $(i = 1, 2, \dots, n)$, через $p(x_i; \Theta)$. *Функцией правдоподобия* дискретной случайной величины X называют функцию аргумента Θ: $L(x_1, x_2, \dots, x_n; \Theta) = p(x_1; \Theta)p(x_2; \Theta) \dots p(x_n; \Theta)$, где $x_1, x_2, \dots, x_n$ – фиксированные числа.

В качестве точечной оценки параметра Θ принимают такое его значение $\Theta^* = \Theta^*(x_1, x_2, \dots, x_n)$, при котором функция правдоподобия достигает максимума. Оценку Θ^* называют *оценкой наибольшего правдоподобия*.

Функции L и $\ln L$ достигают максимума при одном и том же значении Θ, поэтому вместо отыскания максимума функции L ищут (что удобнее) максимум функции $\ln L$.

Логарифмической функцией правдоподобия называют функцию $\ln L$. Как известно, точку максимума функции $\ln L$ аргумента Θ можно искать, например, так:

найти производную

$$(16.19)\ \frac{d\ln L}{d\Theta}$$

приравнять производную нулю и найти критическую точку – корень полученного уравнения (его называют уравнением правдоподобия);

найти вторую производную

$$(16.20)\ \frac{d^2 \ln L}{d\Theta^2}$$

если вторая производная при $\Theta = \Theta^*$ отрицательна, то Θ^* – точка максимума.

Найденную точку максимума Θ^* принимают в качестве оценки наибольшего правдоподобия параметра Θ.

Функция правдоподобия – функция от аргумента Θ; оценка наибольшего правдоподобия – функция от независимых аргументов $x_1, x_2, \dots, x_n$. Оценка наибольшего правдоподобия не всегда совпадает с оценкой, найденной методом моментов.

Пример. Найти методом наибольшего правдоподобия оценку параметра X распределения Пуассона

$$(16.21)\ P_m(X = x_i) = \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!}$$

где m – число произведенных испытаний; x_i – число появлений события в i-м (i = 1, 2, ..., n) опыте (опыт состоит из m испытаний).

Решение. Составим функцию правдоподобия, учитывая, что $\Theta = \lambda$:

$$(16.22)\ L = p(x_1; \lambda) \cdot p(x_2; \lambda) \cdot \ldots \cdot p(x_n; \lambda) = \frac{\lambda^{x_1} \cdot e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} \cdot e^{-\lambda}}{x_2!} \cdot \ldots \cdot \frac{\lambda^{x_n} \cdot e^{-\lambda}}{x_n!} = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{x_1!\, x_2! \ldots x_n!}$$

Найдем логарифмическую функцию правдоподобия:

$$(16.23)\ lnL = \left(\sum x_i\right) ln\lambda - n\lambda - \ln\,(x_1!\, x_2! \ldots x_n!)$$

Найдем первую производную по λ:

$$(16.24)\ \frac{dlnL}{d\lambda} = \frac{\sum x_i}{\lambda} - n$$

Напишем уравнение правдоподобия, для чего приравняем первую производную нулю:

$$(16.25)\ \frac{\sum x_i}{\lambda} - n = 0$$

Найдем критическую точку, для чего решим полученное уравнение относительно λ:

$$(16.26)\ \lambda = \frac{\sum x_i}{n} = \bar{x}_B$$

Найдем вторую производную по λ:

$$(16.27)\ \frac{d^2 lnL}{d\lambda^2} = -\frac{\sum x_i}{\lambda^2}$$

При $\lambda = \bar{x}_B$ вторая производная отрицательна; следовательно, $\lambda = \bar{x}_B$ – точка максимума и, значит, в качестве оценки наибольшего правдоподобия параметра λ распределения Пуассона надо принять выборочную среднюю $\lambda^* = \bar{x}_B$.

Глава семнадцатая. МЕТОДЫ РАСЧЕТА СВОДНЫХ ХАРАКТЕРИСТИК ВЫБОРКИ

Условные варианты

Предположим, что варианты выборки расположены в возрастающем порядке, т. е. в виде вариационного ряда. *Равноотстоящими* называют варианты, которые образуют арифметическую прогрессию с разностью h. *Условными* называют варианты, определяемые равенством $u_i = (x_i - C)/h$, где C – ложный нуль (новое начало отсчета); h – шаг, т. е. разность между любыми двумя соседними первоначальными вариантами (новая единица масштаба). Упрощенные методы расчета сводных характеристик выборки основаны на замене первоначальных вариант условными.

Если вариационный ряд состоит из равноотстоящих вариант с шагом h, то условные варианты есть целые числа.

Обычные, начальные и центральные эмпирические моменты

Для вычисления сводных характеристик выборки удобно пользоваться эмпирическими моментами, определения которых аналогичны определениям соответствующих теоретических моментов. В отличие от теоретических эмпирические моменты вычисляют по данным наблюдений. *Обычным эмпирическим моментом порядка k* называют среднее значение k-х степеней разностей $x_i - C$:

$$(17.1)\ M_k' = \frac{\sum n_i (x_i - C)^k}{n}$$

где x_i – наблюдаемая варианта, n_i – частота варианты, $n = \Sigma n_i$ – объем выборки, C – произвольное постоянное число (ложный нуль).

Начальным эмпирическим моментом порядка k называют обычный момент порядка k при $C = 0$

$$(17.2)\ M_k = \frac{\sum n_i x_i^k}{n}$$

В частности,

$$(17.3)\ M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_B$$

т. е. начальный эмпирический момент первого порядка равен выборочной средней.

Центральным эмпирическим моментом порядка k называют обычный момент порядка k при C = $\bar{x}_B$

$$(17.4)\ m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}$$

В частности,

$$(17.5)\ m_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_B$$

Условные эмпирические моменты

Вычисление центральных моментов требует довольно громоздких вычислений. Чтобы упростить расчеты, заменяют первоначальные варианты условными. *Условным эмпирическим моментом порядка* k называют начальный момент порядка k, вычисленный для условных вариант:

$$(17.6)\ M_k^* = \frac{\sum n_i u_i^k}{n} = \frac{\sum n_i \left(\frac{x_i - C}{h}\right)^k}{n}$$

В частности,

$$(17.7)\ M_1^* = \frac{\sum n_i \left(\frac{x_i - C}{h}\right)}{n} = \frac{1}{h}\left[\frac{\sum n_i x_i}{n} - C\frac{\sum n_i}{n}\right] = \frac{1}{h}(\bar{x}_B - C)$$

Отсюда $\bar{x}_B = M^*_1 h + C$. Таким образом, для того чтобы найти выборочную среднюю, достаточно вычислить условный момент первого порядка, умножить его на h и к результату прибавить ложный нуль C.

Глава восемнадцатая. ЭЛЕМЕНТЫ ТЕОРИИ КОРРЕЛЯЦИИ

Функциональная, статистическая и корреляционная зависимости

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин. Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют *корреляционной*.

Пусть Y – урожай зерна, X – количество удобрений. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т. е. Y не является функцией от X. Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, средний урожай является функцией от количества удобрений, т. е. Y связан с X корреляционной зависимостью.

Выборочный коэффициент корреляции

Выборочный коэффициент корреляции определяется равенством

$$(18.1)\ r_B = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\tilde{\sigma}_x \tilde{\sigma}_y}$$

где x, y – варианты (наблюдавшиеся значения) признаков X и Y; n_{xy} – частота пары вариант (x, y); n – объем выборки (сумма всех частот); $\tilde{\sigma}_x$, $\tilde{\sigma}_y$ – выборочные средние квадратические отклонения; $\bar{x}, \bar{y}$ – выборочные средние.

Предварительные соображения к введению меры любой корреляционной связи

Пусть данные наблюдений над количественными признаками X и Y сведены в корреляционную таблицу. Можно считать, что тем самым наблюдаемые значения Y разбиты на группы; каждая группа содержит те значения Y, которые соответствуют определенному значению X. Например, дана корреляционная таблица:

Y	X	
	8	9
3	4	13
5	6	7
n_x	10	20
$\bar{y}_x$	4,2	3,7

Рис. 9. Корреляционная таблица

К первой группе относятся те 10 значений Y (4 раза наблюдалось $y_1 = 3$ и 6 раз $y_2 = 5$), которые соответствуют $x_1 = 8$. Ко второй группе относятся те 20 значений Y (13 раз наблюдалось $y_1 = 3$ и 7 раз $y_2 = 5$), которые соответствуют $x_1 = 9$.

Условные средние теперь можно назвать групповыми средними: групповая средняя первой группы $\bar{y}_8 = (4*3 + 6*5)/10 = 4,2$; групповая средняя второй группы $\bar{y}_9 = (13*3 + 7*5)/20 = 3,7$. Поскольку все значения признака Y разбиты на группы, можно представить общую дисперсию признака в виде суммы внутригрупповой и межгрупповой дисперсий: $D_{общ} = D_{внгр} + D_{межгр}$. Если Y связан с X функциональной зависимостью, то $D_{межгр} / D_{общ} = 1$. Если Y связан с X корреляционной зависимостью, то $D_{межгр} / D_{общ} < 1$.

Чем связь между признаками ближе к функциональной, тем меньше $D_{внгр}$ следовательно, тем больше приближается $D_{межгр}$ к $D_{общ}$, а значит, отношение $D_{межгр} / D_{общ}$ – к единице. Отсюда ясно, что целесообразно рассматривать в качестве меры тесноты корреляционной зависимости отношение межгрупповой дисперсии к общей, или, что то же, отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению.

Корреляционным отношением η называют отношение $\sigma_{межгр}$ к $\sigma_{общ}$

Корреляционное отношение как мера корреляционной связи

Поскольку в рассуждениях не делалось никаких допущений о форме корреляционной связи, η служит мерой тесноты связи любой, в том числе и линейной, формы. В этом состоит преимущество корреляционного отношения перед коэффициентом корреляции, который оценивает тесноту лишь линейной зависимости. Вместе с тем корреляционное отношение обладает недостатком: оно не позволяет судить, насколько близко расположены точки, найденные по данным наблюдений, к кривой определенного вида, например к параболе, гиперболе и т. д. Эго объясняется тем, что при определении корреляционного отношения форма связи во внимание не принималась.

Глава девятнадцатая. СТАТИСТИЧЕСКАЯ ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Статистическая гипотеза. Нулевая и конкурирующая, простая и сложная гипотезы

Часто необходимо знать закон распределения генеральной совокупности. Если закон распределения неизвестен, но имеются основания предположить, что он имеет определенный вид (назовем его A), выдвигают гипотезу: генеральная совокупность распределена по закону A. Таким образом, в этой гипотезе речь идет о виде предполагаемого распределения.

Возможен случай, когда закон распределения известен, а его параметры неизвестны. Если есть основания предположить, что неизвестный параметр Θ равен определенному значению Θ_0, выдвигают гипотезу: $\Theta = \Theta_0$. Таким образом, в этой гипотезе речь идет о предполагаемой величине параметра одного известного распределения.

Возможны и другие гипотезы: о равенстве параметров двух или нескольких распределений, о независимости выборок и многие другие.

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений. Например, статистическими являются гипотезы: генеральная совокупность распределена по закону Пуассона; дисперсии двух нормальных совокупностей равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения, во второй – о параметрах двух известных распределений. Гипотеза «на Марсе есть жизнь» не является статистической, поскольку в ней не идет речь ни о виде, ни о параметрах распределения.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0. Конкурирующей (альтернативной) называют гипотезу H_1 которая противоречит нулевой. Например, если нулевая гипотеза состоит в предположении, что математическое ожидание a нормального распределения равно 10, то конкурирующая гипотеза, в частности, может состоять в предположении, что $a \neq 10$. Коротко это записывают так: $H_0: a=10; H_1: a \neq 10$.

Различают гипотезы, которые содержат только одно и более одного предположений. *Простой* называют гипотезу, содержащую только одно предположение. Например, если λ – параметр показательного распределения, то гипотеза H_0: λ = 5 – простая. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ известно) – простая.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез. Например, сложная гипотеза H_0: λ > 5 состоит из бесчисленного множества простых вида H_i: $\lambda = b_i$, где b_i – любое число, большее 5. Гипотеза H_0: математическое ожидание нормального распределения равно 3 (σ неизвестно) – сложная.

Ошибки первого и второго рода

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют *статистической*. В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т. е. могут быть допущены ошибки двух родов. *Ошибка первого рода* состоит в том, что будет отвергнута правильная гипотеза. *Ошибка второго рода* состоит в том, что будет принята неправильная гипотеза.

Вероятность совершить ошибку первого рода принято обозначать через α; ее называют уровнем значимости. Наиболее часто уровень значимости принимают равным 0,05 или 0,01. Если, например, принят уровень значимости, равный 0,05, то это означает, что в пяти случаях из ста имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

Статистический критерий проверки нулевой гипотезы. Наблюдаемое значение критерия

Для проверки нулевой гипотезы используют специально подобранную случайную величину, точное или приближенное распределение которой известно. Эту величину обозначают через U или Z, если она распределена нормально, F или υ^2 – по закону Фишера–Снедекора, T – по закону Стьюдента, χ^2 – по закону «хи квадрат» и т. д. Поскольку в этом параграфе вид распределения во внимание приниматься не будет, обозначим эту величину в целях общности через K.

Статистическим критерием (или просто *критерием*) называют случайную величину K, которая служит для проверки нулевой гипотезы. Например, если проверяют гипотезу о равенстве дисперсий двух нормальных генеральных совокупностей, то в качестве критерия K принимают отношение исправленных выборочных дисперсий:

$$(19.1)\ F = \frac{s_1^2}{s_2^2}$$

Эта величина случайная, потому что в различных опытах дисперсии принимают различные, наперед неизвестные значения, и распределена по закону Фишера–Снедекора. Для проверки гипотезы по данным выборок вычисляют частные значения входящих в критерий величин и таким образом получают частное (наблюдаемое) значение критерия.

Наблюдаемым значением $K_{набл}$ называют значение критерия, вычисленное по выборкам. Например, если по двум выборкам найдены исправленные выборочные дисперсии $s_1 = 20$ и $s_2 = 5$, то наблюдаемое значение критерия F

$$(19.2)\ F = \frac{s_1^2}{s_2^2} = \frac{20}{5} = 4$$

Критическая область. Область принятия гипотезы. Критические точки

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, а другая – при которых она принимается.

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

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

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

Поскольку критерий K – одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия гипотезы также являются интервалами и, следовательно, существуют точки, которые их разделяют. *Критическими точками* (границами) $k_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области. *Правосторонней* называют критическую область, определяемую неравенством $K > k_{кр}$, где $k_{кр}$ – положительное число (рис. 10а). *Левосторонней* называют критическую область, определяемую неравенством $K < k_{кр}$, где $k_{кр}$ – отрицательное число (рис. 10б). *Односторонней* называют правостороннюю или левостороннюю критическую область. *Двусторонней* называют критическую область, определяемую неравенствами $K < k_1$, $K > k_2$, где $k_2 > k_1$. В частности, если критические точки симметричны относительно нуля, двусторонняя критическая область определяется неравенствами (в предположении, что $k_{кр} > 0$): $K < -k_{кр}$, $K > k_{кр}$ или равносильным неравенством $|K| > k_{кр}$ (рис. 10в).

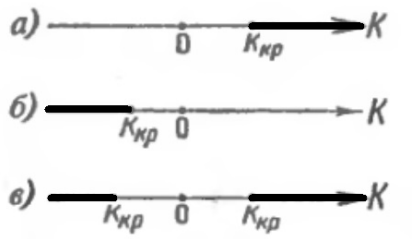

Рис. 10. Односторонние и двусторонняя критические области

Отыскание правосторонней критической области

Как найти критическую область? Для определенности начнем с нахождения правосторонней критической области, которая определяется неравенством $K > k_{кр}$, где $k_{кр} > 0$. Для отыскания правосторонней критической области достаточно найти критическую точку. Для ее нахождения задаются достаточной малой вероятностью – уровнем значимости α. Затем ищут критическую точку $k_{кр}$, исходя из требования, чтобы при условии справедливости нулевой гипотезы вероятность того, что критерий K примет значение, большее $k_{кр}$, была равна принятому уровню значимости: $P(K > k_{кр}) = \alpha$.

Для каждого критерия имеются таблицы[1], по которым и находят критическую точку, удовлетворяющую этому требованию.

Когда критическая точка найдена, вычисляют по данным выборок наблюденное значение критерия и, если окажется, что $K_{набл} > k_{кр}$, то нулевую гипотезу отвергают; если же $K_{набл} < k_{кр}$, то нет оснований отвергнуть нулевую гипотезу.

Наблюдаемое значение критерия может оказаться большим $k_{кр}$ не потому, что нулевая гипотеза ложна, а по другим причинам (малый объем выборки, недостатки методики эксперимента и др.). В

[1] А для многих критериев есть формулы в Excel.

этом случае, отвергнув правильную нулевую гипотезу, совершают ошибку первого рода. Вероятность этой ошибки равна уровню значимости α.

Пусть нулевая гипотеза принята; ошибочно думать, что тем самым она доказана. Действительно, известно, что один пример, подтверждающий справедливость некоторого общего утверждения, еще не доказывает его. Поэтому, более правильно говорить *данные наблюдений согласуются с нулевой гипотезой и, следовательно, не дают оснований ее отвергнуть*.

На практике для большей уверенности принятия гипотезы ее проверяют другими способами или повторяют эксперимент, увеличив объем выборки.

Отвергают гипотезу более категорично, чем принимают. Действительно, известно, что достаточно привести один пример, противоречащий некоторому общему утверждению, чтобы это утверждение отвергнуть. Если оказалось, что наблюдаемое значение критерия принадлежит критической области, то этот факт и служит примером, противоречащим нулевой гипотезе, что позволяет ее отклонить.

Мощность критерия

Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что справедлива конкурирующая гипотеза. Другими словами, мощность критерия есть вероятность того, что нулевая гипотеза будет отвергнута, если верна конкурирующая гипотеза.

Если вероятность ошибки второго рода (принять неправильную гипотезу) равна β, то мощность равна $1 - \beta$.

Если уровень значимости уже выбран, то критическую область следует строить так, чтобы мощность критерия была максимальной. Выполнение этого требования должно обеспечить минимальную ошибку второго рода.

Чем меньше вероятности ошибок первого и второго рода, тем критическая область «лучше». Однако при заданном объеме выборки уменьшить одновременно α и β невозможно; если уменьшить α, то β будет возрастать.

Если α уже выбрано, то, пользуясь теоремой Ю. Неймана и Э. Пирсона, изложенной в более полных курсах, можно построить критическую область, для которой β будет минимальным и, следовательно, мощность критерия максимальной.

Единственный способ одновременного уменьшения вероятностей ошибок первого и второго рода состоит в увеличении объема выборок.

Сравнение двух дисперсий нормальных генеральных совокупностей

Задача сравнения дисперсий возникает, если требуется сравнить точность приборов, инструментов, самих методов измерений и т. д. Очевидно, предпочтительнее тот прибор, инструмент и метод, который обеспечивает наименьшее рассеяние результатов измерений, т. е. наименьшую дисперсию.

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам с объемами, соответственно равными n_1 и n_2, извлеченным из этих совокупностей, найдены исправленные выборочные дисперсии s^2_x и s^2_y. Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: $H_0: D(X) = D(Y)$.

В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий примем отношение большей исправленной дисперсии к меньшей, т. е. случайную величину

$$(19.3)\ F = \frac{S^2_б}{S^2_м}$$

Величина F при справедливости нулевой гипотезы имеет распределение Фишера–Снедекора со степенями свободы $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$, где n_1 – объем выборки, по которой вычислена большая исправленная дисперсия, n_2 – объем выборки, по которой найдена меньшая дисперсия. Напомним, что распределение Фишера–Снедекора зависит только от чисел степеней свободы и не зависит от других параметров.

Критическая область строится в зависимости от вида конкурирующей гипотезы. Если конкурирующая гипотеза $H_1: D(X) \neq D(Y)$, то строят двустороннюю критическую область, исходя из требования, чтобы

вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α.

Наибольшая мощность (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы} достигается тогда, когда вероятность попадания критерия в каждый из двух интервалов критической области равна $\alpha/2$. Если обозначить через F_1 левую границу критической области и через F_2 – правую, то $P(F < F_1) = \alpha/2$, $P(F > F_2) = \alpha/2$.

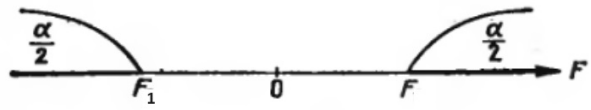

Рис. 11. Критические области при двусторонней конкурирующей гипотезе

Чтобы найти правую критическую точку надо вычислить отношение большей исправленной дисперсии к меньшей (19.3) и по таблице критических точек распределения Фишера–Снедекора по уровню значимости $\alpha/2$ (вдвое меньшем заданного) и числам степеней свободы k_1 и k_2 (k_1 – число степеней свободы большей дисперсии) найти критическую точку $F_{кр}(\alpha/2; k_1, k_2)$.

Уровень значимости α = 0,01

k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6083	6106
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,41	99,42
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,13	27,05
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,452	14,374
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,963	9,888
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,790	7,718
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,538	6,469
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,734	5,667
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	5,178	5,111
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,772	4,706
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,462	4,397
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	4,220	4,155
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	4,025	3,960
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,864	3,800
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,730	3,666
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,616	3,553
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682	3,593	3,519	3,455

Уровень значимости α = 0,05

k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,936	5,912
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,704	4,678
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	4,027	4,000
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,603	3,575
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,313	3,284
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	3,102	3,073
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,943	2,913
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,818	2,788
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,717	2,687
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,635	2,604
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,565	2,534
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,507	2,475
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,456	2,425
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,413	2,381

Рис. 12. Критические точки распределения F Фишера–Снедекора (k_1 – число степеней свободы большей дисперсии, k_2 – число степеней свободы меньшей дисперсии). Для отыскания критической точки в Excel воспользуйтесь формулой $=F.ОБР(1-\alpha; k_1; k_2)$.

Если $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

Далее рассматриваются различные ситуации статистического вывода и различные критерии, в том числе:

- Сравнение исправленной выборочной дисперсии с гипотетической генеральной дисперсией нормальной совокупности
- Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых известны (независимые выборки)
- Сравнение двух средних произвольно распределенных генеральных совокупностей (большие независимые выборки)
- …

ЧАСТЬ ЧЕТВЕРТАЯ. МЕТОД МОНТЕ–КАРЛО. ЦЕПИ МАРКОВА

Глава двадцать первая. МОДЕЛИРОВАНИЕ (РАЗЫГРЫВАНИЕ) СЛУЧАЙНЫ Х ВЕЛИЧИН МЕТОДОМ МОНТЕ-КАРЛО

Сущность метода Монте–Карло состоит в следующем: требуется найти значение a некоторой изучаемой величины. Для этого выбирают такую случайную величину X, математическое ожидание которой равно a: $M(X) = a$.

Практически же поступают так: производят n испытаний, в результате которых получают n возможных значений X; вычисляют их среднее арифметическое $\bar{x}$ и принимают x в качестве оценки (приближенного значения) a^* искомого числа a: $a \approx a^* = \bar{x}$

Разыгрывание дискретной случайной величины

Пусть требуется разыграть дискретную случайную величину X, т. е. получить последовательность ее возможных значений x_i ($i = 1, 2, …, n$), зная закон распределения X:

X	x_1	x_2	…	x_n
P	p_1	p_2	…	p_n

Надо: 1) разбить интервал $(0, 1)$ оси Or на n частичных интервалов: $\Delta_1 – (0; p_1)$, $\Delta_2 – (p_1; p_1 + p_2)$, …, $\Delta_n – (p_1 + p_2 +…+ p_n;1)$; 2) выбрать (например, из таблицы случайных чисел) случайное число r_j. Если r_j попало в частичный интервал Δ_i, то разыгрываемая дискретная случайная величина приняла возможное значение x_1.

18	33	41	69	40	86	91	60	97	43	7	80	54	17	70	12	0	29	18	86
67	62	31	40	74	62	35	4	72	87	18	19	88	77	84	91	74	73	55	96
11	89	97	14	10	25	86	59	17	50	58	55	43	24	63	65	9	50	14	46
84	67	66	84	42	67	60	61	20	70	64	56	85	53	89	87	14	30	71	71
75	57	25	2	18	3	98	60	30	93	82	3	5	83	33	58	21	78	16	43
91	23	44	53	90	58	27	83	4	2	33	51	19	9	36	24	16	70	4	15
74	67	38	39	100	65	14	83	35	77	85	85	82	94	45	21	11	73	89	83
63	43	38	66	88	29	29	80	55	27	61	96	80	92	24	66	57	79	83	24
96	65	31	62	28	45	6	33	45	9	27	47	94	83	18	34	10	4	18	37
74	43	13	81	3	20	59	61	17	22	14	10	83	31	62	28	83	16	22	91
74	45	11	32	25	55	80	34	92	19	39	45	74	56	60	75	42	4	8	56
26	81	12	86	1	12	13	88	85	60	13	25	22	56	30	14	90	83	63	44
95	25	38	62	41	87	54	1	56	49	15	90	58	33	98	78	51	15	61	1
75	34	99	65	5	59	13	75	24	12	73	32	14	39	28	74	14	58	1	24
85	0	23	55	52	10	4	53	35	54	79	50	82	92	4	40	98	42	74	40
76	61	79	71	75	53	35	2	91	37	25	8	17	59	33	47	34	95	92	87
80	87	64	57	17	24	4	86	95	50	14	81	19	69	58	86	21	49	50	53
19	3	64	66	32	36	58	41	88	57	7	85	6	47	32	81	38	93	5	75
59	3	93	25	22	4	75	49	97	16	86	14	85	1	73	88	80	49	28	63
26	52	89	11	82	28	27	23	91	90	41	73	42	5	23	7	23	22	49	91

Рис. 13. Равномерно распределенные случайные числа; число – два первых десятичных знака, например, 18 –> 0,18. В Excel случайное число в диапазоне $(0, 1)$ можно получить формулой =СЛЧИС()

Пример. Разыграть 8 значений дискретной случайной величины X, закон распределения которой задан в виде таблицы

X	3	11	24
p	$0,25$	$0,16$	$0,59$

Разобьем интервал $(0, 1)$ оси Or точками с координатами 0,25; 0,25 + 0,16 = 0,41 на 3 интервала: $\Delta_1 – (0; 0,25)$, $\Delta_2 – (0,24; 0,41)$, $\Delta_3 – (0,41; 1)$. Выпишем из таблицы восемь случайных чисел, например: 0,18; 0,67; 0,11; 0,84; 0,75; 0,91; 0,74; 0,63.

Случайное число r_1 = 0,18 принадлежит частичному интервалу Δ_1, поэтому разыгрываемая дискретная случайная величина приняла возможное значение $x_1 = 3$. Случайное число r_2 = 0,67 принадлежит частичному интервалу Δ_3, поэтому разыгрываемая величина приняла возможное значение x_3 = 24. Аналогично получим остальные возможные значения.

Итак, разыгранные возможные значения X таковы: 3; 24; 3; 24; 24; 24; 24; 24.

Разыгрывание непрерывной случайной величины. Метод обратных функций

Пусть требуется разыграть непрерывную случайную величину X, т. е. получить последовательность ее возможных значений x_i ($i = 1, 2, \ldots, n$), зная функцию распределения $F(x)$.

Теорема. Если r_i – случайное число, то возможное значение x_i разыгрываемой непрерывной случайной величины X с заданной функцией распределения $F(x)$, соответствующее r_i, является корнем уравнения $F(x_i) = r_i$.

Доказательство. Пусть выбрано случайное число r_i ($0 < r_i < 1$). Так как в интервале всех возможных значений X функция распределения $F(x)$ монотонно возрастает от 0 до 1, то в этом интервале существует, причем только одно, такое значение аргумента x_i при котором функция распределения примет значение r_i. Другими словами, уравнение $F(x_i) = r_i$ имеет единственное решение

$$(21.1)\ x_i = F^{-1}(r_i)$$

где F^{-1} – функция, обратная функции $y = F(x)$.

Если решить это уравнение в явном виде не удается, то прибегают к графическим или численным методам.

Пример. Непрерывная случайная величина X распределена по показательному закону, заданному функцией распределения (параметр $\lambda > 0$ известен)

$$(21.2)\ F(x) = 1 - e^{-\lambda x}\ (x > 0)$$

Требуется найти явную формулу для разыгрывания возможных значений X.

Решение. Используя правило (21.1) напишем уравнение

$$(21.3)\ 1 - e^{-\lambda x} = r_i$$

Решим это уравнение относительно x_i:

$$(21.4)\ e^{-\lambda x} = 1 - r_i \text{ или } -\lambda x = \ln(1 - r_i)$$

Отсюда

$$(21.5)\ x_i = -\frac{1}{\lambda}\ln(1 - r_i)$$

Случайное число r_i заключено в интервале (0, 1); следовательно, число $1 - r_i$ также случайное и принадлежит интервалу (0,1). Другими словами, величины R и 1 – R распределены одинаково. Поэтому для отыскания x_i можно воспользоваться более простой формулой

$$(21.6)\ x_i = -\frac{1}{\lambda}\ln r_i$$

Известно, что

$$(21.7)\ F(x) = \int_{-\infty}^{x} f(x)dx$$

В частности,

$$(21.8)\ F(x_i) = \int_{-\infty}^{x_i} f(x)dx$$

Отсюда следует, что если известна плотность вероятности $f(x)$, то для разыгрывания X можно вместо уравнений $F(x_i) = r_i$ решить относительно x_i уравнение

$$(21.9)\quad \int_{-\infty}^{x_i} f(x)dx = r_i$$

Для того чтобы найти возможное значение x_i непрерывной случайной величины X, зная ее плотность вероятности $f(x)$, надо выбрать случайное число r_i и решить относительно x_i уравнение (21.8) или уравнение

$$(21.10)\quad \int_{a}^{x_i} f(x)dx = r_i$$

где a – наименьшее конечное возможное значение X.

Пример. Задана плотность вероятности непрерывной случайной величины X $f(x) = \lambda(1-\lambda x/2)$ в интервале $(0; 2/\lambda)$; вне этого интервала $f(x) = 0$. Требуется найти явную формулу для разыгрывания возможных значений X.

Решение. Напишем в соответствии с правилом (21.9) уравнение

$$(21.11)\ \lambda \int_{0}^{x_i} (1 - \frac{\lambda x}{2})dx = r_i$$

Выполнив интегрирование и решив полученное квадратное уравнение относительно x_i, получим

$$(21.12)\ x_i = 2(1 - \sqrt{1 - r_i})/\lambda$$

Глава двадцать вторая. ПЕРВОНАЧАЛЬНЫЕ СВЕДЕНИЯ О ЦЕПЯХ МАРКОВА

Цепью Маркова называют последовательность испытаний, в каждом из которых появляется только одно из k несовместных событий $A_1, A_2, \ldots, A_k$ полной группы, причем условная вероятность $p_{ij}(s)$ того, что в s-м испытании наступит событие $A_j (j = 1, 2, \ldots, k)$, при условии, что в $(s-1)$-м испытании наступило событие $A_i (i = 1, 2, \ldots. k)$, не зависит от результатов предшествующих испытаний.

Например, если последовательность испытаний образует цепь Маркова и полная группа состоит из четырех несовместных событий A_1, A_2, A_3, A_4, причем известно, что в шестом испытании появилось событие A_2, то условная вероятность того, что в седьмом испытании наступит событие A_4, не зависит от того, какие события появились в первом, втором, ... пятом испытаниях.

Заметим, что независимые испытания являются частным случаем цепи Маркова. Действительно, если испытания независимы, от появление некоторого определенного события в любом испытании не зависит от результатов ранее произведенных испытаний. Отсюда следует, что понятие цепи Маркова является обобщением понятия независимых испытаний.

Далее используется терминология, которая принята при изложении цепей Маркова. Пусть некоторая система в каждый момент времени находится в одном из k состояний: первом, втором, ... k-м. В отдельные моменты времени в результате испытания состояние системы изменяется, т. е. система переходит из одного состояния, например i, в другое, например j. В частности, после испытания система может остаться в том же состоянии («перейти» из состояния i в состояние $j = i$).

Таким образом, события называют состояниями системы, а испытания – изменениями ее состояний. Дадим теперь определение цепи Маркова, используя новую терминологию.

Цепью Маркова называют последовательность испытаний, в каждом из которых система принимает только одно из k состояний полной группы, причем условная вероятность $p_{ij}(s)$ того, что в s-м испытании система будет находиться в состоянии j, при условии, что после $(s-1)$-го испытания она находилась в состоянии i, не зависит от результатов остальных, ранее произведенных испытаний.

Цепью Маркова *с дискретным временем* называют цепь, изменение состояний которой происходит в определенные фиксированные моменты времени.

Цепью Маркова *с непрерывным временем* называют цепь, изменение состояний которой происходит в любые случайные возможные моменты времени.

Однородная цепь Маркова. Переходные вероятности. Матрица перехода

Однородной называют цепь Маркова, если условная вероятность $p_{ij}(s)$ (перехода из состояния i в состояние j) не зависит от номера испытания. Поэтому вместо $p_{ij}(s)$ пишут просто p_{ij}.

Пример. Случайное блуждание. Пусть на прямой Ox в точке с целочисленной координатой $x = n$ находится материальная частица. В определенные моменты времени $t_1, t_2, t_3, \ldots$ частица испытывает толчки. Под действием толчка частица с вероятностью p смещается на единицу вправо и с вероятностью $1 - p$ – на единицу влево. Ясно, что положение (координата) частицы после толчка зависит от того, где находилась частица после непосредственно предшествующего толчка, и не зависит от того, как она двигалась под действием остальных предшествующих толчков.

Таким образом, случайное блуждание – пример однородной цепи Маркова с дискретным временем.

Переходной вероятностью p_{ij} называют условную вероятность того, что из состояния i (в котором система оказалась в результате некоторого испытания, безразлично какого номера) в итоге следующего испытания система перейдет в состояние j.

Таким образом, в обозначении p_{ij} первый индекс указывает номер предшествующего, а второй – номер последующего состояния. Например, p_{11} — вероятность «перехода» из первого состояния в первое; p_{23} — вероятность перехода из второго состояния в третье.

Пусть число состояний конечно и равно k.

Матрицей перехода системы называют матрицу, которая содержит все переходные вероятности этой системы:

$$(22.1)\ \boldsymbol{\mathcal{P}}_1 = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{pmatrix}$$

Так как в каждой строке матрицы помещены вероятности событий (перехода из одного и того же состояния i в любое возможное состояние j), которые образуют полную группу, то сумма вероятностей этих событий равна единице. Другими словами, сумма переходных вероятностей каждой строки матрицы перехода равна единице:

$$(22.2)\ \sum_{j=1}^{k} p_{ij} = 1\ (i = 1, 2, \ldots, k)$$

Приведем пример матрицы перехода системы, которая может находиться в трех состояниях:

$$(22.3)\ \boldsymbol{\mathcal{P}}_1 = \begin{pmatrix} 0{,}5 & 0{,}2 & 0{,}3 \\ 0{,}4 & 0{,}5 & 0{,}1 \\ 0{,}6 & 0{,}3 & 0{,}1 \end{pmatrix}$$

Здесь p_{11} = 0,5 – вероятность перехода из состояния i = 1 в это же состояние j = 1; p_{21} = 0,4 – вероятность перехода из состояния i = 2 в состояние j = 1.

Равенство Маркова

Обозначим через $P_{ij}(n)$ вероятность того, что в результате n шагов (испытаний) система перейдет из состояния i в состояние j. Например, $P_{25}(10)$ – вероятность перехода за 10 шагов из второго состояния в пятое.

Подчеркнем, что при n = 1 получим переходные вероятности $P_{ij}(1) = p_{ij}$.

Поставим перед собой задачу: зная переходные вероятности p_{ij}, найти вероятности $P_{ij}(n)$ перехода системы из состояния i в состояние j за n шагов. С этой целью введем в рассмотрение промежуточное (между i и j) состояние r. Другими словами, будем считать, что из первоначального состояния i за m шагов система перейдет в промежуточное состояние r с вероятностью $P_{ir}(m)$, после чего за оставшиеся $n - m$ шагов из промежуточного состояния r она перейдет в конечное состояние j с вероятностью $P_{rj}(n-m)$.

По формуле полной вероятности

$$(22.4)\ P_{ij}(n) = \sum_{r=1}^{k} P_{ir}(m)P_{rj}(n-m)$$

Эту формулу называют *равенством Маркова*.