Содержание

1. ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1 Распределения случайные величины

Распределения случайной величины бывают дискретными и непрерывными, интегральными и дифференциальными, одномерными и двумерными.

Действительное переменное, которое в зависимости от исхода опыта, т.е. от случая, принимает различные значения, называется случайной величиной. Пусть X – некоторая случайная величина. Функцией распределения F(x) случайной величины X (*X* is random variable)

называется функция:

$$F(x)=P(X<x) \qquad (1)$$

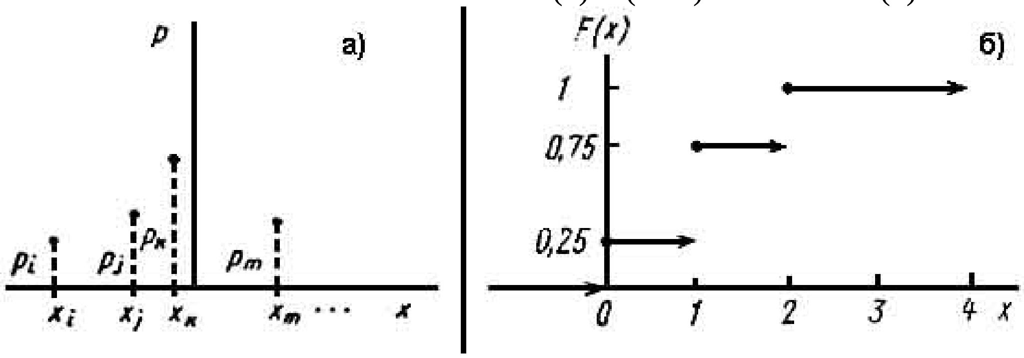


Рис. 1. Распределение вероятностей дискретной случайной величины, f(x) (а) и функция распределения F(x) дискретной случайной величины (б).

Значение функции распределения в точке x_0, таким образом, равно вероятности того, что случайная величина принимает значение, меньшее x_0. В теории вероятностей случайная величина полностью характеризуется своей функцией распределения, т.е. может рассматриваться как заданная, если задана её функция распределения. Функция распределения произвольной случайной величины обладает следующими свойствами:

1) $\lim\limits_{x\to\infty} F(x)=1, \quad \lim\limits_{x\to\infty} F(x)=0$;

2) F(x) является монотонно неубывающей, т.е. при $x_1<x_2$ имеет место $F(x_1)\le F(x_2)$;

3) F(x) непрерывна слева.

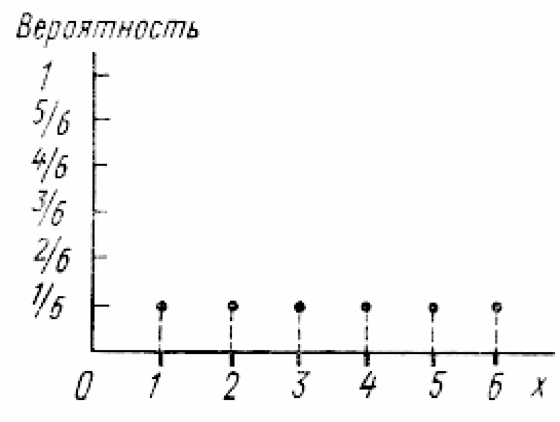


Рис. 2. Распределение вероятностей дискретного равномерного распределения (n=6).

Случайная величина X называется дискретной, если она может принимать только конечное или счётное множество значений. Таким образом, она характеризуется значениями x_1, x_2,….., которые она может принимать, и вероятностями $p_i=P(X=x_i)$, с которыми она принимает эти значения. Вероятности p_i должны удовлетворять условию $\sum_i p_i = 1$.

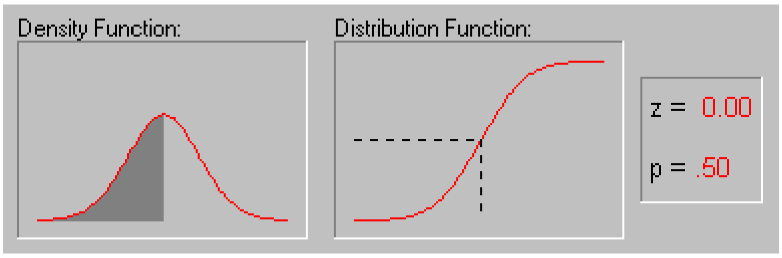


Рис. 3. Дифференциальная и интегральная функции распределения непрерывной величины.

Однозначное отображение множества x_i множество p_i рассматривается как функция вероятности **дискретной** случайной величины.

Для функции распределения дискретной случайной величины имеем:

$$F(x) = \sum_{x_i < x} p_i \qquad (2)$$

Суммирование проводится по всем i, для которых $x_i<x$. Таким образом, F(x) является ступенчатой функцией со скачками высотой p_i в точках x_i.

Примерами дискретных распределений могут служить биномиальное распределение, распределение Пуассона, гипергеометрическое распределение и др.

Случайная величина называется непрерывной, если её функцию распределения (интегральную функцию распределения) можно представить в виде:

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad (3)$$

(F is a cumulative probability distribution function)

Функция f(x) называется плотностью распределения. Так как $\lim_{x\to\infty} F(x) = 1$, то должно выполняться условие:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \qquad (4)$$

(f is a probability density function)

При заданной плотности вероятности, вероятность того, что непрерывная случайная величина попадает в заданный интервал, равна:

$$P(a \le X \le b) = F(b) - F(a) = \int_a^b f(x)dx \qquad (5)$$

1.2 Характеристики распределений

Полная информация о случайной величине дается ее распределением вероятностей (функцией распределения, F, функцией плотности, f). Однако для решения многих задач достаточно знать лишь некоторые числовые характеристики, называемые характеристиками распределения или случайной величины, дающие относительно полное представление о свойствах случайной величины. Важнейшими среди них являются математическое ожидание, дисперсия, среднее квадратическое отклонение, коэффициент вариации, начальные, центральные и основные моменты, коэффициент асимметрии, коэффициент эксцесса, медиана, мода, первая квартиль, третья квартиль, интерквартильный размах и квантили.

1.2.1 Математическое ожидание (среднее арифметическое)

Основной характеристикой положения центра распределения является математическое ожидание (среднее арифметическое).

Рассмотрим случайную величину с числовыми значениями. Часто оказывается полезным связать с этой функцией число – ее «среднее значение» или, как говорят, «среднюю величину» (среднее арифметическое, *arithmetic mean*), «меру центральной тенденции». В качестве «среднего значения» обычно используют математическое ожидание.

Математическое ожидание *(среднее значение) – понятие теории вероятности, важнейшая характеристика распределения значений случайной величины X. В простейшем случае, когда X может принимать лишь конечное число значений* $x_1, x_2, \ldots, x_n$ *с вероятностями* $p_1, p_2, \ldots, p_n$,

Математическим ожиданием величины X называется выражение:

$$EX = x_1p_1 + x_2p_2 + \ldots + x_np_n. \qquad (6)$$

Математическое ожидание (*Expected Value, EV*) является основным понятием в теории вероятностей, которое служит для усредненной оценки некоторого случайного значения. Математическое ожидание похоже на центр тяжести, если считать вероятности значений массами точек. С помощью математического ожидания можно прогнозировать оценку значения некоторого случайного признака при достаточно долгом периоде испытаний. Используя аналогию с центром тяжести, можно сказать, что любое физическое тело, находясь в неопределенном состоянии, будет стремиться принять состояние равновесия (опоры на свой центр тяжести), точно также и среднее значение любой случайной величины при достаточно большом количестве испытаний будет стремиться к своему математическому ожиданию.

В случае, когда вероятности исходов одинаковы ($p_1=p_2=...=p_n=1/n$) формула принимает вид среднего арифметического:

$$EX = \frac{x_1}{n} + \frac{x_2}{n} + \ldots + \frac{x_n}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad (7)$$

Дадим более строгое определение этого параметра.

Математическое ожидание – понятие среднего значения случайной величины в теории вероятностей. Обозначается EX или иногда MX (в русской литературе). В статистике часто используют обозначение μ. Мы в данном обзоре будем обычно обозначать μ_1, имея ввиду, что математическое ожидание – первый начальный момент *(first moment)* от распределения (см. ниже).

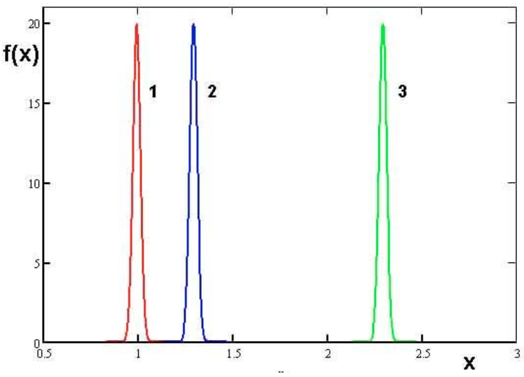


Рис. 4. Распределения с одинаковой дисперсией, но с разными математическими ожиданиями: μ=1 (1); 1,3 (2) и 2,3 (3).

Пусть задано вероятностное пространство (Ω, F, P) и определенная на нём случайная величина X. То есть по определению, X:Ω→R – измеримая функция. Тогда если существует интеграл Лебега от X по пространству Ω, то он называется математическим ожиданием, или средним значением и обозначается EX.

$$EX \equiv \int_{\Omega} X(\omega)P(\omega)d\omega \qquad (8)$$

В дискретном случае:

$$EX = \sum_{\omega\in\Omega} X(\omega)P(\omega), \qquad (9)$$

т.е. математическое ожидание случайной величины – это взвешенная сумма значений случайной величины с весами, равными вероятностям соответствующих элементарных событий.

Свойства математического ожидания.

Если FX(x) – функция распределения случайной величины, то её математическое ожидание задаётся интегралом Лебега – Стилтьеса:

$$EX = \int_{-\infty}^{\infty} x dF_X(x) \qquad (10)$$

Если X – дискретная случайная величина, имеющая распределение $P(X = x_i) = p_i;\quad \sum_{i=1}^{\infty} p_i = 1$, то из определения интеграла Лебега следует, что

$$EX = \sum_{i=1}^{\infty} x_i p_i \qquad (11)$$

Для реальных распределений это выражение записывается как

$$EX = \sum_{i=1}^{i=n} x_i p_i \qquad (12)$$

Математическое ожидание может не существовать, если ряд $\sum_{i=1}^{\infty} x_i p_i$ расходится. Например, если $x_i = n$, $p_n = \frac{1}{n(n+1)}$ $n = 1,2,...$, то $\sum_{n=1}^{\infty} p_n = 1$, а $\sum X = \sum_{n=1}^{\infty} \frac{1}{n+1}$.

Математическое ожидание абсолютно непрерывной случайной величины, распределение которой задаётся плотностью $f_X(x)$, равно

$$EX = \int_{-\infty}^{\infty} x f_X(x)dx \qquad (13)$$

$Ea = a$, если a - постоянная величина;

$$E(aX)=aEX. \qquad (14)$$

Математическое ожидание суммы независимых случайных величин равно сумме их математических ожиданий:

$$E(X+Y)=EX+EY \qquad (15)$$

Пусть X и У – случайные величины, определенные на одном и том же пространстве элементарных событий, a и b – некоторые числа. Тогда

$$M(aX+bY)=aM(X)+bM(Y). \qquad (16)$$

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

$$E(X\cdot Y)=EX\cdot EY \qquad (17)$$

На практике пользуются оценкой математического ожидания, называемой выборочным средним.

Другими характеристиками положения центра распределения являются медиана и мода.

1.2.2 Медиана и мода

Математическое ожидание не всегда является разумной оценкой какой-нибудь случайной величины. Так, для оценки средней заработной платы разумнее использовать понятие медианы, то есть такой величины, что количество людей, получающих меньшую, чем медиана, зарплату и большую, совпадают.

Медиана *– понятие теории вероятности; одна из характеристик распределения значений случайной величины X. Медиана – такое число m, что N принимает с вероятностью ½ как значения большие m, так и меньшие m.*

Медиана в статистике *– значение варьирующего признака, которое делит ряд распределения на две равные части по объёму частот или частей. Сумма абсолютных величин линейных отклонений от медианы минимальна.*

Медианой случайной величины X называется такое число Q_2, что

$$P(X) \le Q_2 \ge \frac{1}{2} \quad и \quad P(X \ge Q_2) \ge \frac{1}{2} \qquad (18)$$

Если X - непрерывная случайная величина, то определение медианы полезно интерпретировать через функцию плотности, как это показано на **Рис. 5.** Таким образом, для непрерывной случайной величины $P(X \le Q_2) = P(X \ge Q_2) = \frac{1}{2}$.

Другими словами, вероятность p_1 того, что случайная величина x окажется меньшей $x_{1/2}$, и вероятность p_2 того, что случайная величина x окажется большей $x_{1/2}$, одинаковы и равны 1/2. Медиана определяется однозначно не для всех распределений.

Если распределение случайной величины симметрично, как, например, в случае нормального распределения, то медиана совпадает с математическим ожиданием. Иногда математическому ожиданию приписывается смысл медианы, что, конечно же, неверно, так как математическое ожидание и медиана для несимметричных распределений, вообще говоря, не совпадают.

Мода *в статистике – величина признака (варианта), чаще всего встречающаяся в совокупности единиц или в вариационном ряду (например, размер одежды, пользующийся наибольшим спросом).*

Модой непрерывной случайной величины называется такое значение X, в котором f(X) достигает своего локального максимума.

Мода есть «центр сгущения» случайной величины в смысле наиболее часто встречающихся (типичных) значений случайной величины. Распределение с одной модой называется унимодальным, а распределение с несколькими модами - мультимодальным. Многовершинность, или многомодальность распределения свидетельствует о существенной неоднородности множества значений исследуемой величины. Для симметричного унимодального распределения мода совпадает с математическим ожиданием, с медианой и с абсциссой центра симметрии графика функции распределения. В этом случае все нечётные центральные моменты равны. Поэтому любой нечётный, тождественно не равный нулю, центральный момент можно рассматривать как характеристику асимметрии распределения. Иногда математическому ожиданию приписывают смысл моды, что, конечно же, неверно, так как математическое ожидание и мода для несимметричных распределений не совпадают.

На **Рис. 5** для несимметричного унимодального распределения показаны все три характеристики положения распределения.

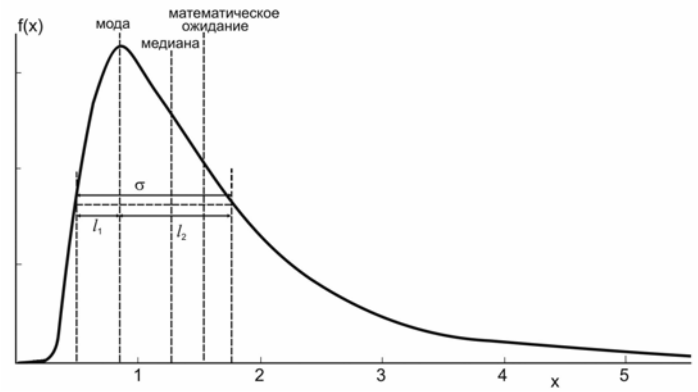


Рис. 5. Мода, медиана и среднее для асимметричного статистического распределения.

1.2.3 Квартили и квантили

Можно ввести еще две характеристики распределения, аналогичные медиане: первую квартиль и третью квартиль.

Первой квартилью распределения случайной величины X называется такое число Q_1 что

$$P(X \le Q_1) \ge \frac{1}{4} \quad и \quad P(X \ge Q_1) \ge \frac{3}{4} \qquad (19)$$

а третьей квартилью распределения случайной величины X называется такое число Q_3, что

$$P(X \le Q_3) \ge \frac{3}{4} \quad и \quad P(X \ge Q_1) \ge \frac{1}{4} \qquad (20)$$

Если X - непрерывная случайная величина, то это определение квартили полезно интерпретировать через функцию плотности в соответствии с **Рис.6** .

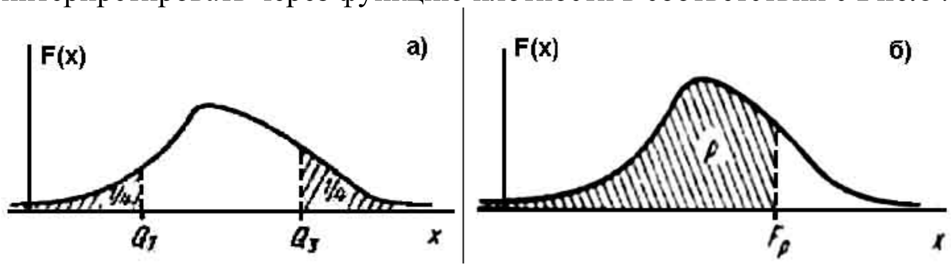


Рис. 6. а) Первая и третья квартили распределения. б) Графическая иллюстрация определения квантили.

Таким образом, для непрерывной случайной величины X $P(X \le Q_1) = P(X \ge Q_3) = \frac{1}{4}$ и $P(X \ge Q_1) = P(X \ge Q_3) = \frac{1}{4}$. Следовательно, вероятность того, что случайная величина примет значение в интервале (Q_1, Q_3), равна $P(Q_1 \le X \le Q_3) = \frac{1}{2}$

Заметим, что длина этого интервала Q_3-Q_1 называется интерквартильным размахом и может служить, аналогично среднеквадратическому отклонению, мерой рассеяния значений случайной величины.

При использовании методов математической статистики, особенно при построении доверительных интервалов и проверки статистических гипотез широко применяется понятие p-квантили.

Квантилью порядка p распределения F(x) называется число C_p, такое, что $F(C_p)$=p.

Кантилью *уровня p или p-квантилью непрерывной случайной величины X с функцией распределения F(x) называется такое значение F_p величины X, что вероятность события $X<F_p$ равна p.*

Для непрерывной случайной величины это определение полезно интерпретировать через функцию плотности**.**

Из определения квантили следует, что медиана есть квантиль порядка 0.5, первая квартиль - квантиль порядка 0.25, а третья квартиль - квантиль порядка 0.75. Для медианы имеем: $P(X<x_{med})=P(X\ge x_{med})=0{,}5$.

Для некоторых, наиболее распространенных в математической статистике распределений созданы таблицы квантилей.

Очень часто, особенно в пакетах прикладных программ по статистической обработке, вместо термина «квантиль» используется термин «процентиль», когда порядок квантили выражается в процентах.

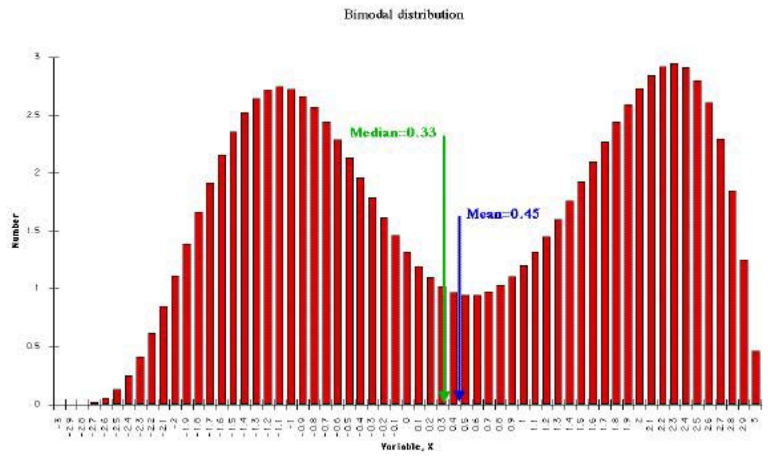


Рис. 7. Среднее и медиана для бимодального распределения.

1.2.4 Дисперсия

(variance of the distribution)

Для описания многих практически важных свойств случайной величины необходимо знание не только ее математического ожидания, но и отклонения возможных ее значений от среднего значения. Для измерения разброса значений случайной величины около среднего значения часто используют такие характеристики как дисперсия, среднее квадратическое отклонение, коэффициент вариации.

Дисперсия *- в теории вероятностей - наиболее употребительная мера отклонения от среднего (мера рассеяния).*

Дисперсия *в математической статистике и теории вероятности – мера рассеяния (отклонения от среднего).*

Дисперсия *– величина, характеризующая степень разброса количественных значений величин статистической выборки (случайных величин) относительно среднего значения для этой выборки. Дисперсия определяется как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания.*

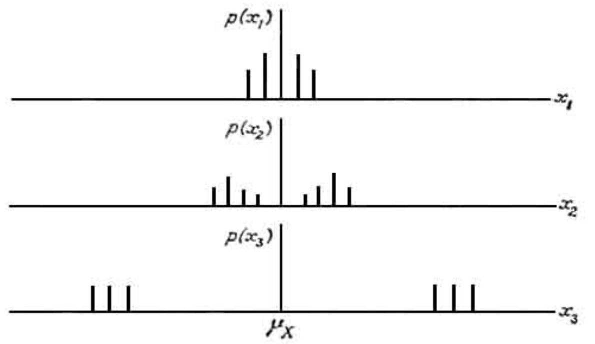


Рис. 8. Распределения случайных величин с одинаковым средним значением, но различной дисперсией.

В статистике дисперсия

$$DX = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} \qquad (21)$$

есть среднее арифметическое из квадратов отклонений наблюдаемых значений (x_1, x_2,…,x_n) случайной величины от их среднего арифметического

$$x = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad (22)$$

В теории вероятности дисперсия случайной величины – математическое ожидание квадрата отклонения случайной величины от её математического ожидания.

Дисперсией случайной величины x называется среднее значение квадрата отклонения случайной величины от её математического ожидания:

$$D_x = \overline{(x - M_x)^2}$$

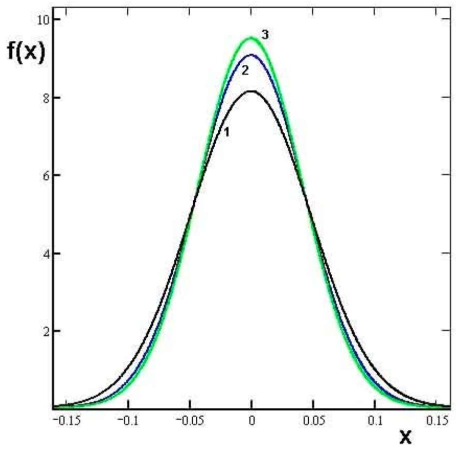


Рис. 9. Распределение с одинаковым математическим ожиданием μ=0 и разными дисперсиями: σ=0,042 (3); 0,044 (2) и 0,049 (1).

Используя вероятности p_i того, что величина x принимает значения x_i, эту формулу можно переписать следующим образом:

$$D_x = \sum_{i=1}^{n} p_i (x_i - M_x)^2. \qquad (23)$$

Среднеквадратическим отклонением случайной величины x называется корень квадратный из дисперсии этой величины *(σ is the standart deviation):*

$$\sigma = \sqrt{D_x} \qquad (24)$$

Дисперсия случайной величины связана с математическим ожиданием квадрата этой случайной величины следующим соотношением:

$$D_x = \sigma^2 = EX^2 - (EX)^2 \qquad (25)$$

Свойства дисперсии.

При умножении случайной величины на постоянную *a* дисперсия увеличивается в a^2 раз: D(aX)= a^2DX; дисперсия всегда неотрицательна: DX≥0. Дисперсия обращается в нуль лишь для вырожденного распределения: если DX=0, то X=Const, и наоборот. Дисперсия не зависит от сдвига случайной величины на постоянную: D(X+a)=DX. Если X_1 и X_2 независимы, то дисперсия их разности равна сумме их дисперсий: $D(X_1-X_2)=DX_1+DX_2$. Для произвольных случайных величин X_1 и X_2 с конечными вторыми моментами имеет место равенство: $D(X_1+X_2)=DX_1+DX_2+2(E(X_1X_2-EX_1EX_2)$. Минимум среднеквадратического отклонения случайной величины X_1 от точек вещественной прямой есть среднеквадратическое отклонение X_1 от своего математического ожидания: $DX_1=E(X-EX)^2 = \min_a(X_1-a)^2$.

$$D(a+X)=DX;\ D(X_1+X_2)=DX_1+DX_2;\ Da=0 \qquad (26)$$

Табл. 1. Основные характеристики некоторых теоретических распределений

Распределение	Математическое ожидание	Дисперсия
Дискретное равномерное (множество возможных значений: $x_1, x_2, \dots, x_n$)	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2$
Распределение Бернулли с параметром *p*	*p*	*p*(1-*p*)
Биномиальное распределение с параметрами *n* и *p*	*np*	*Np*(*1—p*)
Распределение Пуассона с параметром λ	λ	λ
Равномерное распределение на отрезке [*a*,*b*]	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Нормальное распределение с параметрами μ и σ^2 χ^2 - распределение с *n* степенями свободы *t* – распределение с *n* степенями свободы	μ *n* 0	σ^2 2*n* $\frac{n}{n-2}$
F - распределение с *n* и *m* степенями свободы	$\frac{m}{m-2}$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$
Экспоненциальное распределение с параметром *Q*	1/*Q*	$1/Q^2$
Логнормальное распределение с параметрами μ, σ^2	$e^{\mu+\frac{\sigma^2}{2}}$	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$

Понятия математического ожидания и дисперсии обобщаются в системе моментов.

1.2.5 Моменты

Важными числовыми характеристиками случайных величин являются моменты.

Моменты случайной величины имеют аналогию в механике. Так, первый статистический момент равен произведению массы на плечо, а центр тяжести определяется как отношение первого момента к массе. Как среднее значение, так и дисперсия представляют собой моменты, в которых весовой функцией является плотность распределения плотность распределения вероятности. Среднее значение – это начальный момент, а дисперсия – центральный момент. Инвариантные по времени порядка k для одной случайной величины в **Табл.** . Центральный момент третьего порядка служит мерой симметрии распределения случайной величины относительно её среднего значения; центральный момент четвёртого порядка характеризует остроту максимума моды.

Начальные моменты *(first moments).*

Пусть X – дискретная случайная величина с возможными значениями x_1, x_2,… и $p_i=P(X=x_i)$. Число

$$\mu_k = \sum_i X_i^k P_i \qquad (27)$$

называется в случае абсолютной сходимости ряда k-м начальным моментом случайной величины X (или её распределения) (k=1, 2, …).

Особое значение имеет первый начальный момент $\mu_i = \sum_i x_i p_i$, который называется математическим ожиданием X и обозначается MX (иногда EX).

Центральные моменты *(central moments)*.

Число:

$$M_k = \sum_k (x_i - \mu_1)^k P_i \qquad (28)$$

называется центральным k-м моментом X.

Второй центральный момент называется **дисперсией случайной величины** X и обозначается через DX, т.е.

$$DX = \sigma_x^2 = \sum_i (x_i - MX)^2 P_i = M(x - Mx)^2 \qquad (29)$$

Корень квадратный из дисперсии называется разбросом или стандартным отклонением или средним квадратичным отклонением случайной величины и обозначается: $\sigma_x = \sqrt{Dx}$. Среднее квадратическое отклонение является, мерой рассеяния распределения, но измеряется, в отличие от дисперсии, в тех же единицах, которые используют для измерения значений случайной величины. Дисперсия характеризует средний квадрат отклонения случайной величины от своего математического ожидания, т.е. величина σ - мера рассеяния распределения относительно математического ожидания.

Коэффициентом вариации случайной величины X называется число V(X), равное, если MX>0,

$$V(X) = \frac{\sigma(X)}{MX} \qquad (30)$$

Коэффициент вариации является, как и дисперсия, и среднее квадратическое отклонение, мерой рассеяния распределения, но служит для измерения среднего квадратического отклонения в долях математического ожидания.

Замечание. Термин «момент» заимствован из теоретической механики. Известно, что первый момент μ_1, т.е. MX, - это абсцисса центра тяжести массы распределения, а второй центральный момент M_2, т.е. DX, - момент инерции массы относительно перпендикулярной оси, проходящей через центр тяжести MX.

Пусть X – непрерывная случайная величина с плотностью вероятности f(x). Тогда

$$\mu_k = \int_{-\infty}^{\infty} x^k f(x)dx \qquad (31)$$

называется в случае абсолютной сходимости интеграла, k-м **начальным моментом** случайной величины X (k=1, 2, ….)

$$M_k = \int_{-\infty}^{\infty} (x - \mu_1)^k f(x)dx \qquad (32)$$

называется k-м **центральным моментом** случайной величины.

$M_1=0$

$M_2=\sigma^2$ – дисперсия

M_3 – асимметрия

M_4 – эксцесс

Связь между начальными и центральными моментами устанавливается формулами:

$$M_2 = \mu_2 - (\mu_1)^2 \qquad (33)$$

$$M_3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3 \qquad (34)$$

$$M_4 = \mu_4 - 4\mu_1\mu_3 + 6\mu_1^2\mu_2 - 3\mu_1^4 \qquad (35)$$

Замечание. На всякий случай поясню, как выводятся формулы, выражающие центральные моменты через начальные. По определению $M_k = \int_a^b (x - \bar{x})^k p(x)dx$. Разложим множитель $(x - \bar{x})^k$ по биному Ньютона, тогда:

$$M_k = \int_a^b x^k p(x)dx - k\bar{x}\int_a^b x^{k-1} p(x)dx + \frac{k(k-1)}{1\cdot 2}\bar{x}^2 \int_a^b x^{k-2} p(x)dx + \ldots + (-1)^m C_k^m \bar{x}^m \int_a^b x^{k-m} p(x)dx + \ldots$$

$$\ldots + \frac{k(k-1)}{1\cdot 2}\bar{x}^{k-2}(-1)^{k-2}\int_a^b x^2 p(x)dx + (-1)^{k-1}k\bar{x}^{k-1}\int_a^b xp(x)dx + (-1)^k \bar{x}^k \int_a^b p(x)dx \qquad (36)$$

Все интегралы равны начальным моментам: μ_k, μ_{k-1}, …, μ_2, μ_1, μ_0. μ0=1, $\mu_1 = \bar{x}$, поэтому два члена приводятся и, следовательно

$$M_k = \sum_{m=0}^{k-2}(-1)^m C_k^m \bar{x}^m \mu_{k-m} + (-1)^{k-1}(k-1)\bar{x}^k. \qquad (37)$$

В частности: $M_2 = \mu_2 - \bar{x}^2$; $M_3 = \mu_3 - 3\mu_2\bar{x} + 2\bar{x}^3$, $M_4 = \mu_4 - 4\mu_3\bar{x} + 6\mu_2\bar{x}^2 - 3\bar{x}^4$

Статистические моменты допускают простую геометрическую интерпретацию: μ_1 – математическое ожидание, т.е. абсцисса, при которой достигается среднее арифметическое значение из ординат на графике f(x); $M_2=\sigma^2$ – дисперсия – определяет ширину распределения, т.е. расстояние между двумя точками перегиба на графике f(x); M_3 – асимметрия – разность Δ_1-Δ_2, рассчитанная относительно моды распределения – точки, где f(x) достигает максимума, а на графике F(x) наблюдается перегиб; M_4 – эксцесс – радиус кривизны в моде, т.е. островершинность.

Моменты могут быть вычислены и для интегральной плотности распределения:

$$\mu_k = \int_0^\infty F(x)dx^k \qquad (38)$$

Замечание. Моментами можно было бы воспользоваться как суммарными (сводными) числовыми характеристиками случайной величины, но размерность момента порядка k равна размерности X в степени k, что неудобно. Более естественно считать числовыми характеристиками числа $\sqrt[k]{M_k}$, в частности $\sigma = \sqrt{M_2}$; использование именно центральных моментов исключает влияние начало отсчёта величины X, но не исключает влияние масштаба при измерении этой величины.

Рис. 10. Асимметрия *(skewness)*.

По этой причине за моментные числовые характеристики следует принять числа $m_k = \frac{\sqrt[k]{M_k}}{\sigma} = \sqrt[k]{\frac{M_k}{M_2}}$. Числа m_k не зависят от способа измерения величины X, т.е. не зависят ни от выбора начала отсчёта, ни от выбора единиц. Это отвлечённые числа, и их можно использовать для сравнения разнообразных случайных величин между собою. Моментные характеристики дополняют основные характеристики, зависящие и от начала отсчёта и от масштаба: среднее значение $\bar{x}$, медиану m и среднее квадратичное отклонение σ_x. Малость числа m_k при нечётных k может служить указанием, что распределение (плотность вероятности) близко к симметричному относительно центра распределения.

Рис. 11. Эксцесс *(kurtosis)*

Основные моменты *(standardized moments)* – центральные моменты, нормированные на дисперсию

$$r_k = \frac{M_k}{\sigma^k}, \qquad (39)$$

где M_k – k-ый центральный момент, σ - дисперсия.

$r_0=1$;

$$r_1 = \frac{M_1}{\sigma} = \frac{M_1}{\sqrt{M_2}} = 0; \quad r_2 = \frac{M_2}{\sigma^2} = \frac{M_2}{M_2} = 1 \; r_3 = \frac{M_3^3}{\sigma^3} = \frac{M_3}{M_2^{3/2}}; \quad r_4 = \frac{M_4}{\sigma^4} = \frac{M_4}{M_2^2}$$

Параметрические моменты

Важное значение для различных приложений имеет нормированные показатели асимметрии

$$\beta_1 = r_3^2 = \frac{M_3^2}{M_2^3} = \frac{M_3^2}{\sigma^6} \qquad (80)$$

Здесь β_1 – всегда положительное число (т.е. правосторонняя и левосторонняя асимметрии совпадают).

и эксцесса:

$$\beta_2 = r_4 = \frac{M_4}{M_2^2} = \frac{M_4}{\sigma^4} \qquad (81).$$

Мера β_1 характеризует соотношение асимметрии к мере рассеяния и может быть использована для сравнения асимметрии двух распределений, имеющих различный масштаб. Все симметричные распределения будут иметь нулевой коэффициент асимметрии. Аналогично, мера β_2 характеризует островершинность распределений, нормированных на математическое ожидание и дисперсию. Эксцесс характеризует поведение кривой плотности в окрестности модального значения, её островершинность.

Нормировка на дисперсию введена для получения безразмерной характеристики, не зависящей от выбора физической единицы измерения значений случайной величины X.

Замечание 1. Здесь приведены формулы для параметрических моментов в записи их автора – Пирсона. В последнее время обычно под первым параметрическим моментом понимают не β_1, а $r_4 = \pm\sqrt{\beta_1} = \gamma_1$, обозначая его как k_1 (или какой иной буквой, но с цифрой 1), что создаёт изрядную путаницу.

k_1 может быть как положительным, так и отрицательным числом, т.е. правосторонняя и левосторонняя асимметрии не совпадают

Иногда модернизируют, заменяя β_2 на $k_2=\beta_2$-3= r_4-3=γ_2 (см. систему SATISTICA),поскольку тогда для основного статистического распределения – нормального (гауссового) распределения, β_1=0 и β_2=0 и карта Пирсона имеет в качестве начала координат нормальное распределение.

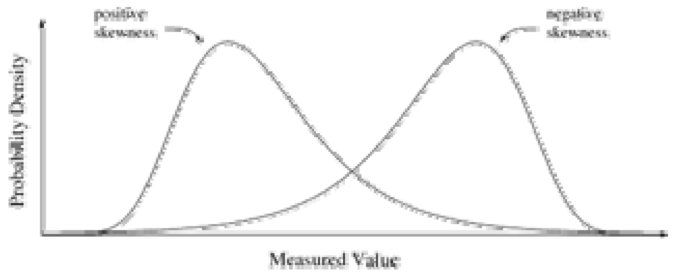


Рис. 12. Положительная и отрицательная асимметрия.

В данной работе мы под симметрией будем понимать параметр

$\gamma_1=r_3=\pm\beta_1^{1/2}$; (82)

а под эксцессом

$$\gamma_2=\beta_2\text{-}3=r_4\text{-}3 \qquad (83)$$

и именно на них будем строить диффузионную карту.

Замечание 2. Программа STATISTICA понимает под симметрией

$$Skewness = \frac{nM_3}{(n-1)(n-2)\sigma^3}, \qquad (83)$$

где $M_3=\Sigma(X_i=\text{СреднееX})^3$; σ^3 – стандартное отклонение в третьей степени, т.е. $M_3^{3/2}$; n – число наблюдений.

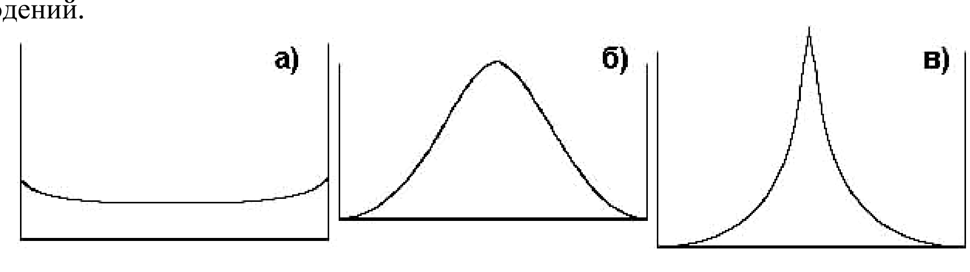


Рис. 13. Виды эксцесса, $k = \dfrac{n\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4}{\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} - 3$, где $\overline{X}$ - среднее: отрицательный эксцесс, нормальный и положительный.

Под эксцессом STATISTICA понимает

$$Kurtosis = \frac{n\cdot(n+1)\cdot M_4 - 3\cdot M_2^2\cdot(n-1)}{(n-1)\cdot(n-2)\cdot(n-3)\cdot\sigma^4} \qquad (84)$$

Программа MATHCAD рассчитывает симметрию по формуле

$$skew(A,B,C,...)=\frac{m\cdot n}{(m\cdot n-1)\cdot(m\cdot n-2)}\sum_{i=0}^{m-1}\sum_{j=0}^{\infty}\left(\frac{M_{i,j}-mean(M)}{Stdev(M)}\right)^3$$ где M множество созданное из аргументов функций A, B, C… Стандартное отклонение M не должно быть равным нулю.

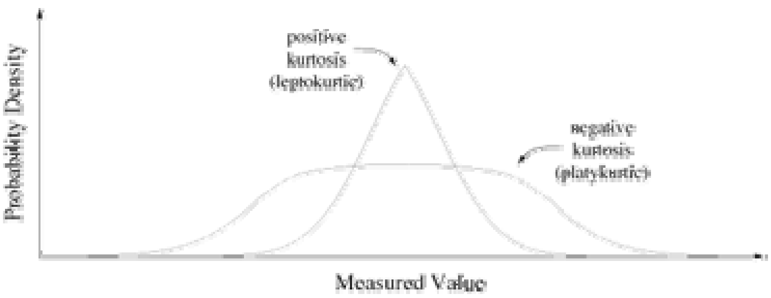


Рис. 14 . Положительный и отрицательный эксцесс.

y = skewness(X) функция предназначена для расчета точечной оценки коэффициента асимметрии y выборки X. Если X задана как вектор, то точечная оценка коэффициента асимметрии рассчитывается по всем его элементам. Для выборки определенной в виде матрицы точечная оценка коэффициента асимметрии рассчитывается для каждого столбца X. Расчет проводится по формуле

$y_j=\frac{E\left(X_j-\overline{X}_j\right)}{\sigma_j^3}$, где $\overline{X}_j$ - среднее арифметическое значение выборки X, σ_j- точечная оценка среднего квадратического отклонения выборки X, E(t) - наиболее вероятная оценка параметра t. Коэффициент асимметрии выборки является мерой смещенности распределения относительно среднего арифметического значения.

Отрицательный коэффициент асимметрии соответствует распределению смещенному влево относительно среднего значения. Положительный коэффициент асимметрии соответствует распределению смещенному вправо относительно среднего значения. Для нормального закона, или любого другого симметричного распределения, коэффициент асимметрии равен нулю.

y = skewness(X,flag) функция позволяет рассчитать несмещенную (flag=0) и смещенную (flag=1, значение по умолчанию) точечную оценку коэффициента асимметрии y выборки X. Величина смещения выборочного коэффициента асимметрии зависит от объема выборки. Если X является выборкой из генеральной совокупности, для получения несмещенной точечной оценки коэффициента асимметрии flag должен быть равен 0.

Эксцесс в MATHADе рассчитывают по формуле

$$kurt(A,B,C...)=\left[\frac{m\cdot n\cdot(m\cdot n+1)}{(m\cdot n-1)\cdot(m\cdot n-2)\cdot(m\cdot n-3)}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left(\frac{M_{i,j}-mean(M)}{Stdev(M)}\right)^4\right]-\frac{3\cdot(m\cdot n-1)^2}{(m\cdot n-2)\cdot(m\cdot n-3)} \quad (85)$$

Если kurt(M)>0, то распределение более островершинное, чем нормальное (у него kurt=0), а если kurt(M)<0, то распределение более плоское, чем нормальное.

Используются две функции: k = kurtosis(X) и k = kurtosis(X,flag)

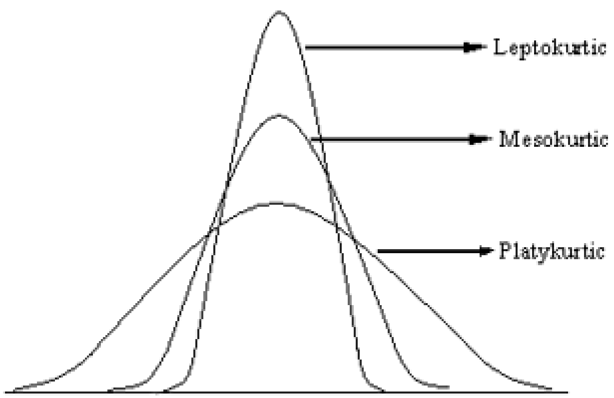


Рис. 15. Примеры Leptokurtic распределения: Лапласа, Логистическое, пример Mesokurtic распределения: нормальное распределение, пример Platykurtic распределения: однородное (непрерывное или дискретное).

k = kurtosis(X) функция предназначена для расчета точечной оценки коэффициента эксцесса k выборки X. Если X задана как вектор, то точечная оценка коэффициента эксцесса рассчитывается по всем его элементам. Для выборки определенной в виде матрицы точечная оценка коэффициента эксцесса рассчитывается для каждого столбца X. Расчет точечной оценки коэффициента эксцесса выборки выполняется по формуле

$k_j=\frac{E\left(X_j-\overline{X}_j\right)^4}{\sigma_j^4}$, где $\overline{X}_j$ - среднее арифметическое значение выборки X_j, σ_j - точечная оценка среднего квадратического отклонения выборки X_j, σ_j - наиболее вероятная оценка параметра t. Коэффициент эксцесса показывает насколько выборка X по наклону кривой функции плотности вероятности соответствует нормальному закону. *Для нормального закона определённый* ***здесь*** *коэффициент эксцесса равен 3.* Законы распределения с более острой вершиной, чем у нормального имеют коэффициент эксцесса более 3 и с менее острой вершиной - менее 3.

Примечание: как мы уже неоднократно указывали, в отечественной литературе коэффициент эксцесса определяется по формуле $k_j = \frac{E\left(X_j - \overline{X}_j\right)^4}{\sigma_j^4} - 3.$

k = kurtosis(X,flag) функция позволяет рассчитать несмещенную (flag=0) и смещенную (flag=1, значение по умолчанию) точечную оценку коэффициента эксцесса k выборки X. Величина смещения выборочного коэффициента эксцесса зависит от объема выборки. Если X является выборкой из генеральной совокупности, для получения несмещенной точечной оценки коэффициента эксцесса flag должен быть равен 0.

Рассмотрим два параметрических момента несколько подробнее.

Коэффициент асимметрии задает степень асимметричности плотности вероятности относительно оси, проходящей через ее центр тяжести. Коэффициент асимметрии (skewness) - безразмерная величина - определяется третьим центральным моментом распределения. В любом симметричном распределении с нулевым математическим ожиданием, например, нормальным, все нечетные моменты, в том числе и третий, равны нулю, поэтому коэффициент асимметрии тоже равен нулю. Асимметрия или коэффициент асимметрии (термин был впервые введен Пирсоном, 1895) является мерой несимметричности распределения. Если асимметрия отчетливо отличается от 0, распределение асимметричное, плотность нормального распределения симметрична относительно среднего. Асимметрия распределения с длинным правым хвостом положительна. Если распределение имеет длинный левый хвост, то его асимметрия отрицательна.

Степень сглаженности плотности вероятности в окрестности главного максимума задается еще одной величиной - **коэффициентом эксцесса** (kurtosis, γ_2). Он показывает, насколько острую вершину имеет плотность вероятности по сравнению с нормальным распределением: коэффициент асимметрии положителен, если правый хвост распределения длиннее левого, и отрицателен в противном случае. Для нормального закона γ_2=0. Если γ_2>0, то распределение имеет острый пик, если γ_2<0 (минимальное значение γ_4= –2), то распределение имеет плосковершинную форму по сравнению с рассмотренным ниже нормальным распределением, для которого γ_4= 0.

Кроме рассмотренных величин используются также так называемые **абсолютные моменты:**

Абсолютный начальный момент: $\beta_k = M[|X|^k]$.

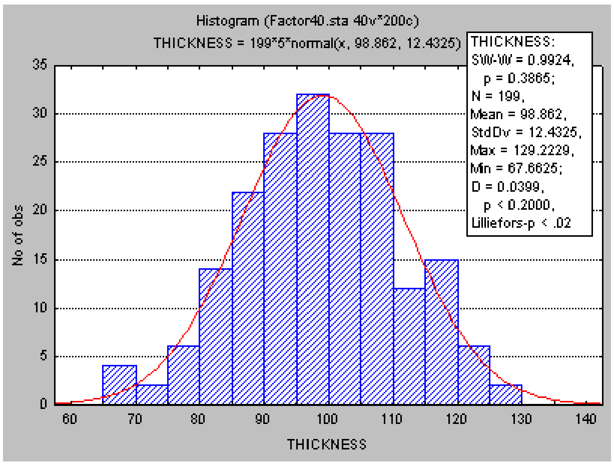
Абсолютный центральный момент: $\nu_k = M[|X - m_x|^k]$.

Абсолютный центральный момент первого порядка называется средним арифметическим отклонением.

Более точную информацию о форме распределения можно получить с помощью критериев нормальности (например, критерия Колмогорова-Смирнова или W критерия Шапиро-Уилка). Однако ни один из этих критериев не может заменить визуальную проверку с помощью гистограммы (графика, показывающего частоту попаданий значений переменной в отдельные интервалы).

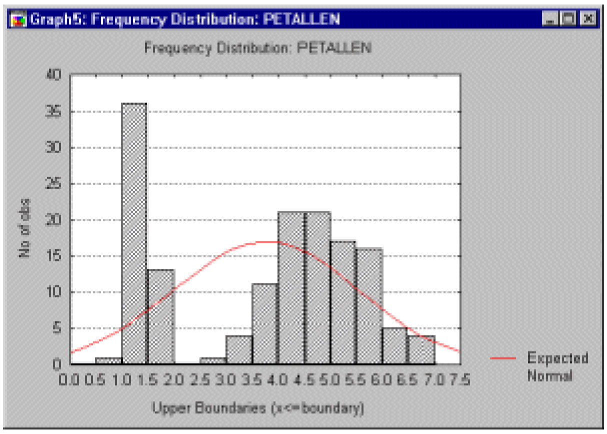
1.2.6 Гистограмма

2M гистограммы являются графическими представлениями распределения частот выбранных переменных, на которых для каждого интервала (класса) рисуется столбец, высота которого пропорциональна частоте класса.



Система STATISTICA предлагает несколько типов 2M гистограмм: простые для одиночных переменных, с двойной осью y для изображения двух переменных/гистограмм на одном графическе с общей шкалой частот, составные для изображения составных переменных/гистограмм на одной графической плоскости с общей шкалой частот и гистограмму висячих столбцов для «подвешивания» столбцов частот переменной к нормальной кривой. Последний тип является наглядным критерием проверки на нормальность распределения, определить области распределения, где возникают расхождения между наблюдаемыми и ожидаемыми нормальными частотами.

Гистограмма позволяет "на глаз" оценить нормальность эмпирического распределения. На гистограмму также накладывается кривая нормального распределения. Гистограмма позволяет качественно оценить различные характеристики распределения. Например, на ней можно увидеть, что распределение бимодально (имеет два пика). Это может быть вызвано, например, тем, что выборка неоднородна, возможно, извлечена из двух разных популяций, каждая из которых более или менее нормальна. В таких ситуациях, чтобы понять природу наблюдаемых переменных, можно попытаться найти качественный способ разделения выборки на две части.



Среднее арифметическое является самым распространенным из набора величин, оценивающих расположение или центральную тенденцию тела данных распределения. Однако среднее арифметическое является не единственным доступным измерением центральной тенденции, и зачастую не самым лучшим. Среднее арифметическое обычно оказывается плохим выбором, когда распределение имеет широкие хвосты. Если при исследовании распределения с очень широкими хвостами вы случайным образом будете выбирать точки данных для расчета среднего, то, проделав это несколько раз подряд, увидите, что средние арифметические, полученные таким способом, заметно отличаются друг от друга. Еще одной важной величиной, определяющей расположение распределения, является медиана (median). Медиана описывает среднее значение, когда данные расположены по порядку в соответствии с их величиной. Медиана делит распределение вероятности на две половины таким образом, что площадь под кривой одной половины равна площади под кривой другой половины. В некоторых случаях медиана лучше задает центральную тенденцию, чем среднее арифметическое. В отличие от среднего арифметического медиана не искажается крайними случайными значениями. Более того, медиану можно рассчитать даже для распределения, в котором все значения выше заданной ячейки попадают в определенную ячейку.

Третьей величиной, определяющей центральную тенденцию, является мода (mode) — наиболее часто повторяющееся событие (или значение данных). Мода — это пик кривой распределения. В некоторых распределениях нет моды, а иногда есть более чем одна мода. Как и медиана, мода в некоторых случаях может лучше всего описывать центральную тенденцию. Мода никак не зависит от крайних случайных значений, и ее можно рассчитать быстрее, чем среднее арифметическое или медиану. Мы увидели, что медиана делит распределение на две равные части. Таким же образом распределение можно разделить тремя квартилями (quartiles), чтобы получить четыре области равного размера или вероятности, или девятью децилями (deciles), чтобы получить десять областей равного размера или вероятности, или 99 перцентилями (percentiles) (чтобы получить 100 областей равного размера или вероятности), 50-й перцентиль является медианой и вместе с 25-м и 75-м перцентилями дает нам квартили. И наконец, еще один термин, с которым вы должны познакомиться, — это квантиль (quantile). Квантиль — это некоторое число N-1, которое делит общее поле данных на N равных частей.

Выше мы обсудили среднее арифметическое, которое измеряет центральную тенденцию распределения. Есть и другие виды средних, они реже встречаются, но в определенных случаях также могут оказаться предпочтительнее. Одно из них — это среднее геометрическое (geometric mean), которое является корнем степени N из произведения значений, соответствующих точкам распределения.

$$G = \frac{\prod_{i=1}^{N} X_i}{N}, \qquad (86)$$

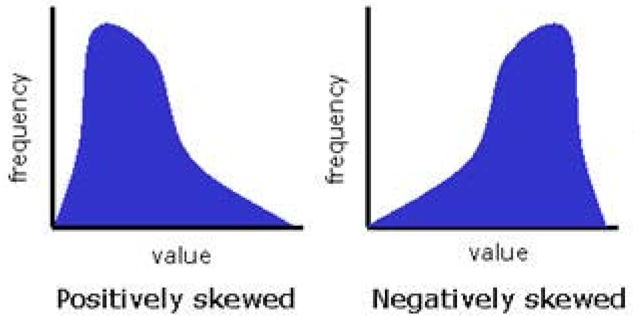
где G = среднее геометрическое; X = значение, соответствующее точке i; N = общее число точек данных в распределении.

Среднее геометрическое не может быть рассчитано, если хотя бы одна из переменных меньше или равна нулю.

Еще одним видом среднего является среднее гармоническое (harmonic mean). Это обратное значение от среднего обратных значений точек данных.

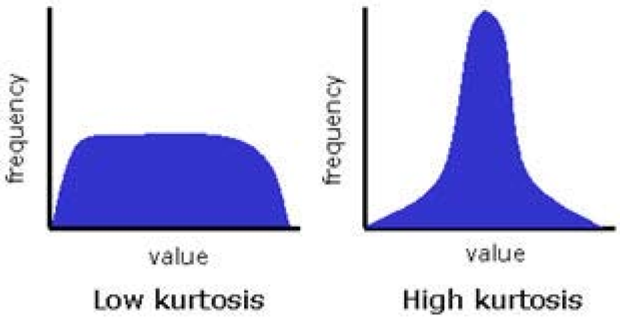
$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}}, \qquad (87)$$

где H = среднее гармоническое; X = значение, соответствующее точке i; N = общее число точек данных в распределении.



$$\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

Рис. 16. Положительная и отрицательная асимметрия распределения.



$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4$$

Рис. 17. Распределения с различным эксцессом.

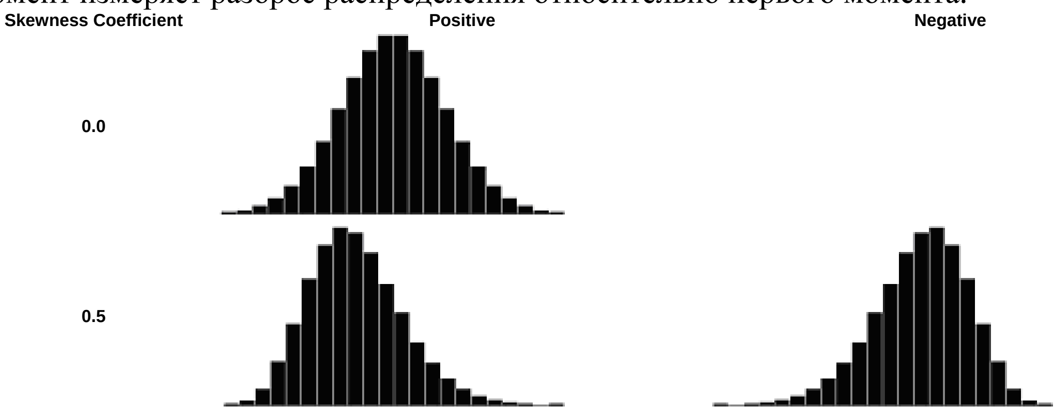
Последней величиной, определяющей центральную тенденцию, является среднее квадратическое (quadratic mean), или среднеквадратический корень (root mean square).

$$R^2 = \frac{1}{N} \sum_{i=1}^{n} X_i^2, \qquad (88)$$

где R = среднеквадратический корень; X = значение, соответствующее точке i; N = общее число точек данных в распределении.

Среднее арифметическое (A) всегда больше или равно среднему геометрическому (G), а среднее геометрическое всегда больше или равно среднему гармоническому (H): $H<G<A$, где H – среднее гармоническое, G = среднее геометрическое; A = среднее арифметическое.

Центральное значение, или расположение распределения, — первое, что надо знать о группе данных. Следующая величина, которая представляет интерес, — это изменчивость данных, или «ширина» относительно центрального значения. Изменчивость точек данных относительно центральной тенденции называется вторым моментом распределения. Следовательно, второй момент измеряет разброс распределения относительно первого момента.



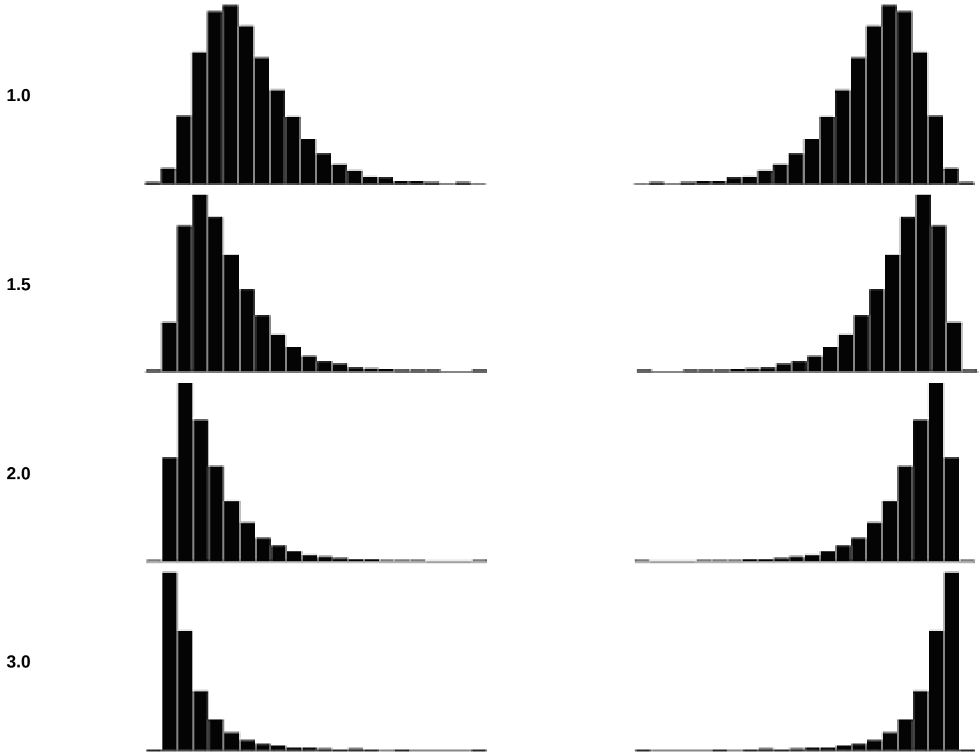


Рис. 18. Распределения с различным значением коэффициента асимметрии.

Как и в случае с центральной тенденцией, существует много способов измерения разброса. Далее мы рассмотрим семь из них, начиная с наименее распространенных вариантов и заканчивая самыми распространенными.

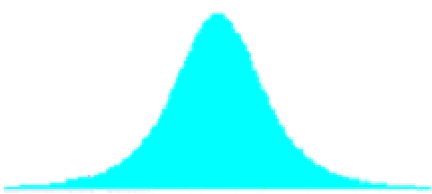
Широта (range) распределения — это просто разность между самым высоким и самым низким значением распределения. Таким же образом широта перцентиля 10-90 является разностью между 90-й и 10-й точками. Эти первые две величины измеряют разброс по крайним точкам. Остальные пять измеряют отклонение от центральной тенденции (т.е. измеряют половину разброса).

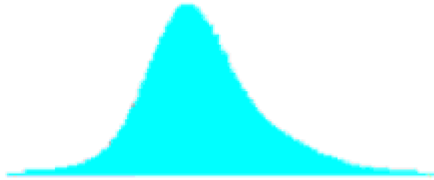
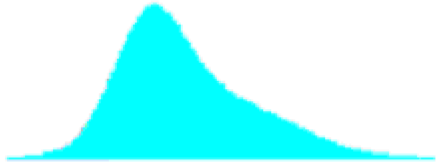
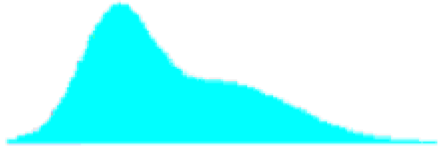
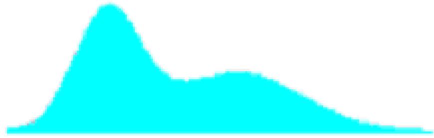
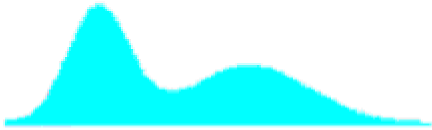
Семи-интерквартильная широта (sem-interquartile range), или квартальное отклонение (quartile deviation), равна половине расстояния между первым и третьим квартилями (25-й и 75-й перцентили). В отличие от широты перцентиля 10-90, здесь широта делится на два. Полуширина (half-width) является наиболее распространенным способом измерения разброса. Сначала надо найти высоту распределения в его пике (моде), затем найти точку в середине высоты и провести через нее горизонтальную линию перпендикулярно вертикальной линии. Горизонтальная линия пересечет кривую распределения в одной точке слева и в одной точке справа. Расстояние между этими двумя точками называется полушириной. Среднее абсолютное отклонение (mean absolute deviation), или просто среднее отклонение, является средним арифметическим абсолютных значений разности значения каждой точки и среднего арифметического значений всех точек. Другими словами (что и следует из названия), это среднее расстояние, на которое значение точки данных удалено от среднего. В математических терминах:

$$M = \frac{1}{N}\sum_{i=1}^{N} ABC(X_i - A), \qquad (89)$$

где M = среднее абсолютное отклонение; N = общее число точек данных; X. = значение, соответствующее точке i; A = среднее арифметическое значений точек данных; ABS() = функция абсолютного значения.

Табл. 2. Распределения с различными средними и стандартными отклонениями

Отношение стандартных отклонений	Среднее в единицах взвешенного стандартного отклонения	Форма комбинированного распределения
2.0	0.00	

2.0	0.67	
2.0	1.33	
2.0	2.00	
2.0	2.67	
2.0	3.33	

Уравнение ( ) дает нам совокупное среднее абсолютное отклонение. Вам следует знать, что можно рассчитать среднее абсолютное отклонение по выборке. Для расчета среднего абсолютного отклонения выборки замените 1/N в уравнении (3.06) на 1/(N - 1). Используйте эту версию, когда расчеты ведутся не по всей совокупности данных, а по некоторой выборке.

Самыми распространенными величинами для измерения разброса являются дисперсия и стандартное отклонение. Как и в случае со средним абсолютным отклонением, их можно рассчитать для всей совокупности и для выборки. Далее показана версия для всей совокупности данных, которую можно легко переделать в выборочную версию, заменив l/NHal/(N-l). Дисперсия (variance) чем-то напоминает среднее абсолютное отклонение, но при расчете дисперсии каждая разность значения точки данных и среднего значения возводится в квадрат. В результате, нам не надо брать абсолютное значение каждой разности, так как мы автоматически получаем положительный результат, независимо от того, была эта разность отрицательной или положительной. Кроме того, так как в квадрат возводится каждая из этих величин, крайние выпадающие значения оказывают большее влияние на дисперсию, а не на среднее абсолютное отклонение. В математических терминах:

$$V = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - A\right)^2 , \qquad (90)$$

где V = дисперсия; N = общее число точек данных; X. = значение, соответствующее точке i; A = среднее арифметическое значений точек данных.

Стандартное отклонение (standard deviation) тесно связано с дисперсией (и, следовательно, со средним абсолютным отклонением). Стандартное отклонение является квадратным корнем дисперсии.

Рис. 19. Асимметрия (а) и эксцесс (б)

Третий момент распределения называется асимметрией (skewness), и он описывает асимметричность распределения относительно среднего значения (**Рис.19**). В то время как первые два момента распределения имеют размерные величины (то есть те же единицы измерения, что и измеряемые параметры), асимметрия определяется таким способом, что получается безразмерной.

Это просто число, которое описывает форму распределения. Положительное значение асимметрии означает, что хвосты больше с положительной стороны распределения, и наоборот. Совершенно симметричное распределение имеет нулевую асимметрию.

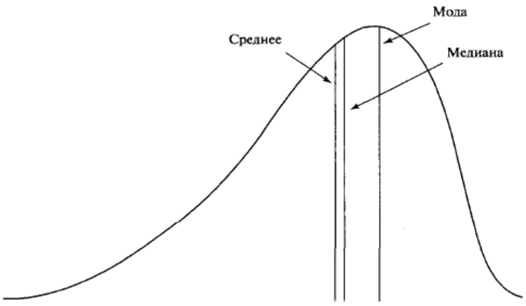


Рис. 20. Асимметричное распределение

В симметричном распределении среднее, медиана и мода имеют одинаковое значение. Однако когда распределение имеет ненулевое значение асимметрии, оно может принять вид, показанный на **Рис. 20.** Для асимметричного распределения (любого распределения с ненулевой асимметрией) верно равенство:

Среднее - Мода = 3 * (Среднее - Медиана)

Есть много способов для расчета асимметрии, и они часто дают различные ответы.
Ниже мы рассмотрим несколько вариантов:

$$S = \begin{cases} \dfrac{Среднее - Мода}{Стандартное\ отклонение} \\ \dfrac{Среднее - Медиана}{Стандартное\ отклонение} \end{cases}$$

Эти выражения дают нам первый и второй коэффициенты асимметрии Пирсона. Асимметрия также часто определяется следующим образом: $S = \dfrac{1}{N}\sum_{i=1}^{N}\left(\dfrac{X_i - A}{D}\right)^3$, где S = асимметрия; N = общее число точек данных; X = значение, соответствующее точке i; A = среднее арифметическое значений точек данных; D = стандартное отклонение значений точек данных.

И наконец, четвертый момент распределения, эксцесс (kurtosis), измеряет, насколько у распределения плоская или острая форма (по сравнению с нормальным распределением). Как и асимметрия, это безразмерная величина. Кривая, менее остроконечная, чем нормальная, имеет эксцесс отрицательный, а кривая, более остроконечная, чем нормальная, имеет эксцесс положительный. Когда пик кривой такой же, как и у кривой нормального распределения, эксцесс равен нулю, и мы будем говорить, что это распределение с нормальным эксцессом.

Как и предыдущие моменты, эксцесс имеет несколько способов расчета. Наиболее распространенными являются: $K = \dfrac{Q}{P}$, где K = эксцесс; Q == семи-интерквартильная широта; P = широта перцентиля 10-90.

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{X - A_i}{D}\right)^4 - 3, \qquad (91)$$

где K = эксцесс; N = общее число точек данных; X = значение, соответствующее точке i; A = среднее арифметическое значений точек данных; D = стандартное отклонение значений точек данных.

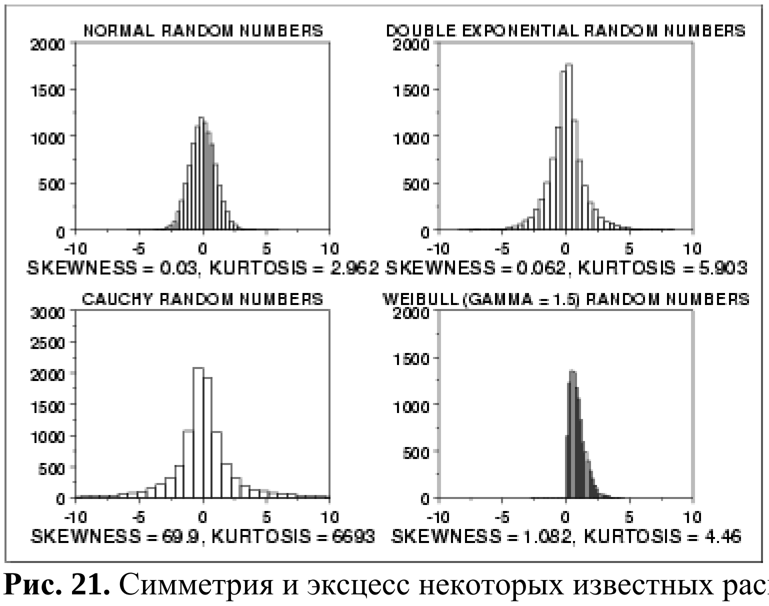


Рис. 21. Симметрия и эксцесс некоторых известных распределений.

1.2.7 Моменты многомерной случайной величины

Совокупность $X_1, X_2,\ldots,X_n$ случайных величин называется n-мерным случайным вектором, который может быть охарактеризован своей n-мерной функцией распределения:

$$F(x_1,\ldots,x_n) = P(X_1 < x_1,\ldots,X_n < x_n) \qquad (92)$$

Функцию $F(x_1, x_2, \ldots x_n)$ часто также называют кратко распределением вектора $(X_1, X_2,\ldots,X_n)$ или совместным распределением величин $X_1, X_2,\ldots,X_n$.Функцию распределения непрерывного вектора можно представить в виде:

$$F(x_1,x_2,\ldots,x_n) = \int_{-\infty}^{x_1}\ldots\int_{-\infty}^{x_n} f(t_1,\ldots,t_n)dt_1\ldots dt_n \qquad (93)$$

При этом должно выполняться условие:

$$\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(x_1,\ldots,x_n)dx_1\ldots dx_n = 1 \qquad (94)$$

Если случайный вектор $(X_1, X_2,\ldots,X_n)$ дискретен, то числа

$$\mu_j = \sum_{i=1,\ldots,i_n} x_{ij}^{(j)} P_{i,\ldots,i_n} \quad (j=1,\ldots,n) \qquad (95)$$

называются первыми начальными моментами $X_1, X_2,\ldots,X_n$. В непрерывном случае первые начальные моменты определяются формулой:

$$\mu_j = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} x_j f(x_1,\ldots,x_n)dx \quad (j=1,\ldots,n) \qquad (96)$$

μ_j являются математическими ожиданиями отдельных компонент: $\mu_j = MX_j$. Вторые начальные моменты μ_{ij} и вторые центральные моменты M_{jk} определяются следующим образом:

- дискретный случай:

$$\mu_{jk} = \sum_{i_1\ldots i_n} x_{ij}^{(j)} x_{ij}^{(k)} P_{i,\ldots,i_n} \qquad (97)$$

$$M_{jk} = \sum_{i_1,\ldots,i_n} \left(x_{ij}^{(j)} - \mu_i\right)\left(x_{ij}^{(k)} - \mu_k\right) P_{i_k,\ldots,i_n} \qquad (98)$$

- непрерывный случай

$$\mu_{jk} = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} x_j x_k f(x_1,\ldots,x_n)dx_1,\ldots,dx_n \qquad (99)$$

$$M_{ik} = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} (x_j - \mu_j)(x_k - \mu_k) f(x_i,\ldots,x_n)dx_1,\ldots,dx_n \qquad (100)$$

Имеем: $\mu_{jk} = M(X_j X_k)$

$$M_{ik} = M\left[(X_j - MX_j)(X_k - M(X_k)\right] = M(X_j X_k) - MX_j MX_k = \mu_{jk} - \mu_j \mu_k \qquad (101)$$

Величины M_{ij} равны дисперсиям отдельных компонент:

$$M_{ij} = DX_j = \sigma_j^2 \qquad (102)$$

Величина $M_{jk} = Cov(X_i, X_k)$ называется ковариацией (корреляционным моментом) случайных величин X_j, X_k обозначается $Cov(X_j,X_k)$. Матрица $\left\| M_{jk} \right\|_{j=1,\ldots,n}$ называется матрицей ковариации. Параметр

$$\rho_{jk} = \frac{Cov(X_j, X_k)}{\sqrt{DX_j DX_k}} = \frac{Cov(X_j, X_k)}{\sigma_j \sigma_k} \qquad (103)$$

называется коэффициентом корреляции между X_j и X_k. Он лежит между -1 и +1. Две корреляции или их ковариация равна нулю.

<u>Пример.</u> В случае нормального распределения при n=2:

$$f(X_1,X_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(X_1 - a_1)}{\sigma_1^2} - 2\rho\frac{(X_1 - a_1)(X_2 - a_2)}{\sigma_1\sigma_2} + \frac{(X_2 - a_2)^2}{\sigma_2^2}\right]\right\}$$

(104)

тогда значения параметров $a_1=MX_1$, $a_2=MX_2$; $\sigma_1=\sqrt{DX_1}$, $\sigma_2=\sqrt{DX_2}$, $\rho=\dfrac{Cov(X_1,X_2)}{\sqrt{DX_1DX_2}}$ (коэффициент корреляции между X_1 и X_2).

1.3 Функции от случайных величин

Пусть дан непрерывный случайный вектор (X, Y), f(x,y) – его плотность. Требуется найти распределение случайных величин X+Y, X·Y, X/Y.

Сумма X+Y есть также непрерывная случайная величина и имеет плотность

$$f(z)=\int_{-\infty}^{\infty} f(x,z-x)dx \qquad (105)$$

Если X и Y независимы, так что $f(x,y)=f_1(x)f_2(x)$, то $f(z)=\int_{-\infty}^{\infty} f_1(x)f(z-x)dx$. Следовательно, плотность суммы есть свёртка плотностей отдельных слагаемых.

Пример. Пусть (X,Y) распределён нормально, а f(x,y) задана формулой (*). Тогда для плотности X и Y получается:

$$f(z)=\frac{1}{\sqrt{2\pi\left(\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2\right)}}\exp\left\{-\frac{1}{2}\frac{[z-(a_1+a_2)]^2}{\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2}\right\} \qquad (106)$$

Следовательно, Z=X+Y снова нормально распределена с параметрами $\sqrt{\sigma_1^2+2\rho\sigma_1\sigma_2+\sigma_2^2}$, a_1+a_2.

В случае независимости имеет место обратное: если сумма двух независимых случайных величин нормально распределена, то и отдельные слагаемые распределены нормально.

Рассмотрим теперь произведение. Пусть (X,Y) – случайный вектор и Z=X·Y. Тогда для плотности имеем:

$$f(z)=\int_{-\infty}^{\infty} f\left(x,\frac{z}{x}\right)\frac{1}{|x|}dx \qquad (107)$$

Перейдём теперь к отношению. Пусть (X, Y) – случайный вектор; отношение X/Y есть некоторая случайная непрерывная функция с функцией плотности:

$$f(z)=\int_{0}^{\infty} xf(zx,x)dx-\int_{-\infty}^{0} xf(zx,x)dx \qquad (108)$$

Табл. 3. Общая сводка формул для функции случайной величины, распределённой по нормальному закону.

Функция случайной величины	Плотность распределения
$\omega=a+bx$, a и b постоянные	$P(\omega)=\left\lvert\frac{1}{b}\right\rvert f\left(\frac{\omega-a}{b}\right)$
$\omega=\frac{1}{x}$	$P(\omega)=\frac{1}{\omega^2}f\left(\frac{1}{\omega}\right)$
$\omega=e^x$	$P(\omega)=\left\lvert\frac{1}{\omega}\right\rvert f(\ln\omega)$
$\omega=\ln x$	$P(\omega)=e^{\omega}f\left(e^{\omega}\right)$
$\omega=x^2$	$P(\omega)=\frac{1}{2\sqrt{\omega}}\left[f(+\sqrt{\omega})+f\left(-\sqrt{\omega}\right)\right]$
$\omega=xy$	$P(\omega)=\int\left\lvert\frac{1}{z}\right\rvert f(\omega z)g(z)dz=\int\left\lvert\frac{z}{\omega^2}\right\rvert f(z)g(\frac{z}{\omega})dz$
$\omega=x+y$	$P(\omega)=\int_{z_1}^{z_2} f(z)g(\omega-z)dz$ - при увеличении числа слагаемых P(ω) стремится к нормальному распределению
$\omega=\frac{x}{y}$	$P(\omega)=\int(z)f(\omega z)g(z)dz=\int\left\lvert\frac{z}{\omega^2}\right\rvert f(z)g\left(\frac{z}{\omega}\right)dz$

Пример. Найти дисперсию функции одной или большего числа случайных величин не находя распределения самой функции.

Пусть $X_1, X_2, \ldots, X_n$ случайных величин с математическим ожиданиями $E(x_1), E(x_2), \ldots, E(x_n)$, дисперсиями $\sigma^2(x_1), \sigma^2(x_2), \ldots, \sigma^2(x_n)$ и ковариацией $Cov(X_i, X_j)$. Тогда дисперсия линейной функции этих случайных величин имеет вид:

$$\sigma^2\left(\sum_{j=1}^{n} b_j X_j\right) = \sum_{j=1}^{n} b_j^2 \sigma^2(X_j) + 2\sum_i \sum_j b_i b_j Cov(X_i, X_j), \qquad (109)$$

где $b_1, b_2, \ldots, b_n$ – постоянные. Для двух случайных величин X_1 и X_2:

$$\sigma^2(b_1 X_1 + b_2 X_2) = b_1^2 \sigma^2(X_1) + b_2^2 \sigma^2(X_2) + 2 b_1 b_2 Cov(X_1, X_2) \qquad (110)$$

Следствия: если n величин некоррелированы, то

$$\sigma^2\left[\sum_{j=1}^{n} b_j X_j\right] = \sum_{j=1}^{n} b_j^2 \sigma^2(X_j) \qquad (111)$$

Если C константа: $\sigma^2(C)=0$; $\sigma^2(Cx)=C^2\sigma^2(x)$

Существует следующее утверждение: если сумма независимых нормально распределённых величин также имеет нормальное распределение с математическим ожиданием равным сумме отдельных математических ожиданий и дисперсией, равной сумме отдельных дисперсий. Для больших выборок, взятых из совокупностей, распределённых по любому закону, сумма случайных величин (т.е. распределение среднего) имеет нормальное распределение.

Рассмотрим теперь общий случай функциональной связи между распределениями. Пусть соотношение между характеристикой системы Z и случайными параметрами отдельных компонентов $X_1, X_2, \ldots, X_n$ задано функцией $Z=h(X_1, X_2, \ldots, X_n)$. Пусть $E(X_i)$ – среднее значение или математическое ожидание случайного параметра i-го компонента, а $M_k(X_i)$ - её k-й центральный момент. Найти оценки $E(Z)$ и $M_k(Z)$ для k=2, 3 и 4 на основе: 1) данных о случайных параметрах компонентов, с помощью которых можно получить оценки и для i=1,2,…,n; 2) знания структуры системы $h(X_1, X_2, \ldots, X_n)$.

Метод состоит в разложении функции $h(X_1, X_2, \ldots, X_n)$ в многомерный ряд Тейлора относительно $E(X_1), E(X_2), \ldots, E(X_n)$, т.е. точки, в которой случайный параметр каждого компонента принимает значение, равное его математическому ожиданию. Сохраняя члены до второго порядка включительно,

$$E(Z) = h[E(X_1), E(X_2, \ldots, E(X_n)] + \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2 \overline{h}}{\partial X_i^{\,2}} \sigma^2(X_i) \qquad (112)$$

где $\dfrac{\partial^2 \overline{h}}{\partial X_i^2}$ - значение производной $\dfrac{\partial^2 h}{\partial X_i^2}$ в точке $E(X_r)$, т.е. вместо X_r (r=1, 2, …,n) подставлена $E(X_r)$.

Пример. Если $h(Z)=X_1 \cdot X_2^2 \cdot X_3^3$, то $\dfrac{\partial^2 \overline{h}}{\partial X_3^2} = [E(X_1)] \cdot [E(X_2)]^2 \cdot [bE(X_3)]$ и т.д.

$$\sigma^2(z) = \sum_{i=1}^{n}\left(\frac{\partial \overline{h}}{\partial X_i}\right)^2 \sigma^2(X_i) + \sum_{i=1}^{n}\left(\frac{\partial \overline{h}}{\partial X_i}\right)\left(\frac{\partial^2 \overline{h}}{\partial X_i^{\,2}}\right) M_3(X_3)$$

если брать разложение до третьего члена включительно. Обычно пользуются выражением:

$$\sigma^2(Z) = \sum_{i=1}^{n}\left(\frac{\partial h}{\partial X_i}\right)\sigma^2(X_i)$$

$$M_3(Z) = \sum_{i=1}^{n}\left(\frac{\partial h}{\partial X_i}\right) M_3(X_i)$$

$$M_4(Z) = \sum_{i=1}^{n}\left(\frac{\partial \overline{h}}{\partial X_i}\right) M_4(X_i) + b\sum_i^{n}\sum_j^{n}\left(\frac{\partial \overline{h}}{\partial X_j}\right)^2 \sigma^2(X_i)\sigma^2(X_j) \text{ (Здесь i>j)}$$

1.4 Характеристические функции

Под характеристической функцией $\psi(t)$ случайной величины X понимают математическое ожидание случайной величины e^{itX}:

$$\Psi(t) = M\left(e^{itX}\right), \qquad (113)$$

где t – действительный параметр.

Если F(x) – функция распределения X, то

$$\Psi(t)=\int_{-\infty}^{\infty} e^{itx}dF(x) \qquad (114)$$

В случае дискретного распределения

$$\Psi(t)=\sum_{k=0}^{\infty} e^{itx_k}P_k \qquad (115)$$

(ряд Фурье с коэффициентами P_k). В случае непрерывного распределения

$$\Psi(t)=\int_{=\infty}^{\infty} e^{itx}f(x)dx \qquad (116)$$

(интеграл Фурье).

Пример 1. Пусть X подчиняется закону Пуассона с параметром λ. Тогда характеристической функцией будет:

$$\Psi(t)=\sum_{k=0}^{\infty} e^{ikt}\frac{\lambda^k}{k!}e^{-\lambda}=e^{\lambda\left(e^{it}-1\right)}$$

Пример 2. Пусть X равномерно распределена на (-a,+a). Тогда

$$\Psi(t)=\int_{-a}^{a} e^{itx}\frac{1}{2a}dx=\frac{Sinat}{at}$$

Пример 3. Если $X\in N(x,a,\sigma)$, то характеристической функцией будет

$$\Psi(t)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{itx-\frac{1}{2}\frac{(x-a)^2}{\sigma^2}}dx=e^{iat-\frac{\sigma^2t^2}{2}}$$

Используя характеристические функции можно легко вычислить моменты.

Если случайная величина X обладает моментом порядка n, то характеристическая функция X n раз дифференцируема по t и при k<n

$$\Psi^{(k)}(0)=i^kMx^k=i^k\mu_k \qquad (117)$$

Пример. Пусть X нормально распределена с параметрами a и σ, $\Psi(t)=e^{iat-\frac{\sigma^2t}{2}}$, тогда $i\mu_1=\psi'(0)=ia$ и $\mu_2=\psi''(0)=-\sigma^2-a^2$. Следовательно, $Mx=\mu_1=a$ и $Dx=\mu_2-\mu_1^2=\sigma^2+a^2-a^2=\sigma^2$. Для приложений имеет значение тот факт, что характеристическая функция суммы независимых случайных величин равна произведению их характеристических функций. Если X_1 и X_2 непрерывны, то плотность суммы есть свёртка обеих плотностей. Это свойство ни что иное, как теорема о свёртке преобразования Фурье.

Пример теоремы о свёртке. Пусть X распределена биномиально с параметрами n и p. Требуется найти характеристическую функцию случайной величины X. Имеем: $X_1+X_2+\ldots+X_n$, где $X_j=\begin{cases}0, \textit{если A не осуществляется в } j-\textit{м опыте}\\ 1, \textit{если A осуществляется в } j-\textit{м опыте}\end{cases}$. По условию X- независимые случайные величины, следовательно, согласно теореме о свёртке, имеем $\Psi_X(t)=\prod_{j=1}^{n}\Psi_{X_j}(t)$, причём $\Psi_{X_j}(t)=Me^{itX_j}=e^{it0}q+e^{it1}p=q+pe^{it}$. Таким образом,

$$\Psi_X(t)=\left(q+pe^{it}\right)^n \qquad (118)$$

Формула обращения. Пусть F(x) – функция распределения, а ψ(t) – характеристическая функция случайной величины X. Если x_1 и x_2 – точки непрерывности F(x), то

$$F(x_2)-F(x_1)=\frac{1}{2\pi}\lim_{e\to\infty}\int_{-e}^{+e}\frac{e^{-itx_1}-e^{-itx_2}}{it}\Psi(t)dt \qquad (119)$$

Если X непрерывна, а f(x) - плотность F(x), то (35) упрощается:

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\Psi(t)dt \qquad (120)$$

Таким образом, плотность распределения получается обратным преобразованием Фурье.

Пример. Пусть две независимые случайные величины нормально распределены $X\in N(x,a_1,\sigma_1)$, $Y(y,a_2,\sigma_2)$ требуется найти распределение для X+Y. Т.к.

$$\Psi_X(t)=e^{ia_1t-\sigma_1^2\frac{t}{2}}$$

$$\Psi_Y(t) = e^{ia_2 t - \sigma_2^2 \frac{t}{2}}$$

то, согласно теореме о свёртке

$$\Psi_{X+Y}(t) = \Psi_X(t)\Psi_Y(t) = e^{i(a_1+a_2)t - \frac{(\sigma_1^2+\sigma_2^2)}{2}}$$

Вследствие теоремы единственности, единственное распределение, имеющее эту функцию распределения, есть N(X+Y), a_1+b_1, $(\sigma_1^2+\sigma_2^2)^{1/2}$. Таким образом, X+Y снова распределена нормально с параметрами: a= a_1+b_1и $(\sigma_1^2+\sigma_2^2)^{1/2}$.

Можно также ввести **производящую функцию моментов** (путем использования преобразования Лапласа):

$$M_X'(L) = E\left(e^{xs}\right) = \begin{cases} \int\limits_{-\infty}^{\infty} e^{sx} f(x)dx & - \ \textit{непрерывна} \\ \sum\limits_{r=0}^{\infty} e^{rs} P_r & - \ \textit{дискретна} \end{cases} \tag{121}$$

где s – любое комплексное число, для которого приведённые интегралы или ряды сходятся абсолютно.

Производящая функция моментов сводит набор начальных моментов к единственному выражению. Действительно, разложим экспоненты в ряд:

$$M_X' = E\left[1 + xs + \frac{(xs)^2}{2!} + \frac{(xs)^3}{3!} + \ldots\right] = 1 + \mu_1 s + \mu_2 \frac{s^2}{2!} + \mu_3 \frac{s^3}{3!} + \ldots, \tag{122}$$

так что μ_k будет просто коэффициентом при $\frac{s^k}{k!}$. Иными словами:

$\mu_k \equiv k! \cdot$коэффициент при s_k.

Другой способ нахождения μ_k из производящей функции моментов (ПФМ) MX'(s) состоит в том, что эту функцию дифференцируют k раз. В результате имеем:

$$\frac{\partial^k M_X'(s)}{\partial s^k} = \int\limits_{-\infty}^{\infty} x^k e^{xs} f(x)dx \tag{123}$$

Отсюда следует:

$$\frac{\partial^k M_X'(0)}{\partial s^k} = \int\limits_{-\infty}^{\infty} x^k f(x)dx = \mu_k \tag{124}$$

Аналогично, производящая функция центральных моментов случайной величины X:

$$M_X(s) = E\left[e^{(X-\mu_1)s}\right] \tag{125}$$

Разлагая в ряд, получаем:

$$M_X(s) = \varepsilon\left[1 + (x-\mu_1)s + (x-\mu_1)^2 \frac{s^2}{2!} + \ldots\right] = 1 + 0 + M_2 \frac{s^2}{2!} + M_3 \frac{s^3}{3!} + \ldots \tag{126}$$

так что M_k – коэффициент при $s\frac{k}{k!}$. Заметим, что $M(s) = e^{-\mu s} M'(s)$ и $M'(s) = e^{\mu s} M(s)$. Наконец, если две ПФМ одинаковы, то одинаковы и исходные распределения вероятности.

Данные методы имеют много общего с использованием преобразования Лапласа:

$$L(s) = \int\limits_{0}^{\infty} e^{-xs} f(x)dx \tag{127}$$

Таким образом, найдя моменты, подставив их в ПФМ и применяя обратное преобразование, получим искомую плотность вероятности.

Примеры нахождения моментов от ПФМ.

Биноминальное распределение.

$$M'(s) = \sum_0^{\infty} e^{rs} P_r = \sum_0^{\infty} C_N^r \left(pe^s\right)^r (1-p)^{N-r} = \left(pe^t + 1 - p\right)^N,$$

$$\frac{\partial M'(s)}{\partial s} = Npe^s\left(pe^s + 1 - p\right)^{N-1},$$

$$\mu_1 = \frac{\partial M'(0)}{\partial t} = Np.$$

Распределение Пуассона:

$$\mu'(s) = e^{-\mu}\sum_{0}^{\infty}\frac{\left(\mu e^{s}\right)^{2}}{r!} = e^{-\mu}e^{\mu e^{s}} = e^{\mu\left(e^{s}-1\right)},$$

$$\frac{\partial M'(s)}{\partial s} = \mu e^{s}e^{\mu}\left(e^{s}-1\right)$$, т.е. $\mu_1=\mu$

$$M(s) = e^{-\mu s}M'(s) = e^{\mu\left(1+s+\frac{s^2}{2!}+\frac{s^3}{3!}+...-t-1\right)} = e^{-\mu\left(\frac{s^2}{2!}+\frac{s^3}{3!}+...\right)} = 1+\mu\left(\frac{s^2}{2!}+\frac{s^3}{3!}+...\right)+\mu^2\frac{1}{2}\left(\frac{s^2}{2!}+\frac{s^3}{3!}+...\right)^2+...$$

Коэффициент при $s^2/2$ равен μ. Значит $M_2=\sigma^2=\mu$.

Нормальное распределение:

$$M'(s) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{xs}e^{-(x-\mu)^2/2\sigma^2}dx = e^{\mu s+\sigma^2 s^2/2} = e^{\mu s}e^{\sigma^2 s^2/2} = \left[1+\mu s+\mu^2\frac{s^2}{2!}+...\right]\left[1+\sigma^2\frac{s^2}{2!}+...\right]$$ Коэффициент при s $\mu_1=\mu$, коэффициент при $s^2/2!$ $\mu_2=\mu^2+\sigma^2$. Следовательно $M_2=\mu_2-(\mu_1)^2=\sigma^2$.

$$M(s) = e^{-\sigma^2\frac{s^2}{2!}} = \sum_{k=0}^{\infty}\left(\sigma^2\frac{s^2}{2!}\right)^k\frac{1}{k!} = \sum_{n=0}^{\infty}\mu_n\frac{s^n}{n!}$$

Таким образом, $M_{2k} = \frac{(2k)!}{2^{kn!}}\sigma^{2k}$ и $M_{2n+1}=0$/

Распределение y=ax+b, где a и b – постоянные величины.

$$M'_y(s) = e^{\mu as+\sigma^2a^2s^2/2+bs} = e^{(a\mu+b)s+a^2\sigma^2s^2/2}$$

Следовательно y распределено по нормальному закону со средним aμ+b и дисперсией $a^2\sigma^2$.

Гамма-распределение.

$f(x) = \frac{1}{\alpha!\beta^{\alpha+1}}x^{\alpha}e^{-\frac{x}{\beta}};\quad 0\le x\le\infty$, при α=0, $f(x) = \frac{1}{\beta}e^{-\frac{x}{\Delta}}$. ПФМ:

$$M'(s) = \frac{1}{\alpha!\beta^{\alpha+1}}\int_{-\infty}^{\infty}x^{\alpha}e^{-\frac{x}{\beta}}e^{xs}dx$$

Преобразуем это выражение к константе, умноженной на интеграл от плотности. Тогда M'(s) будет просто равна этой константе. Подставим $-\frac{x}{\beta}+xs = -x\frac{1-\beta s}{\beta}$ и, кроме того, $\beta^{\alpha+1} = \left[\frac{\beta}{(1-\beta s)}\right]^{\alpha+1}(1-\beta s)^{\alpha+1}$. В таком случае

$$M'(s) = \frac{1}{(1-\beta s)^{\alpha+1}}\int_{-\infty}^{\infty}\left(\frac{1}{\alpha!}\right)\left(\frac{1-\beta s}{\beta}\right)x^{\alpha}e^{-\frac{x(1-\beta s)}{\beta}}dx = \frac{1}{(1-\beta s)^{\alpha+1}}\int_{-\infty}^{\infty}f\left(x,\alpha,\frac{\beta}{1-\beta s}\right)dx$$

Следовательно,

$$M'(s) = (1-\beta s)^{-\alpha-1}$$