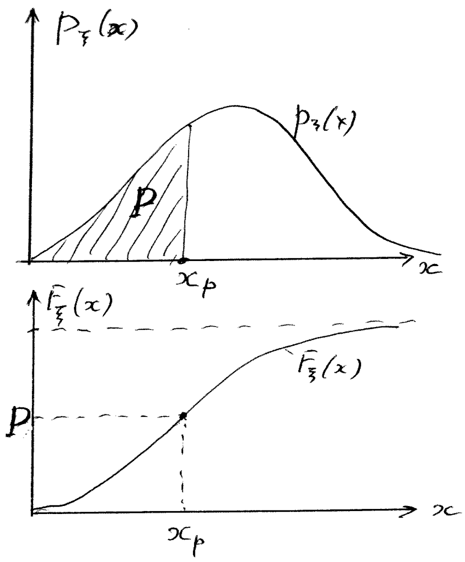
Лекция 5

5.3. Метод порядковых статистик.

В статистике, кроме системы моментов, в качестве числовых характеристик распределений широко используются числовые характеристики, называемые *квантилями*.

Определение. *Значение x_p случайной величины ξ называется **p-квантилью**, если*



$$P\{\xi < x_p\} = p,$$

где x_p — это корень уравнения

$F_\xi(x_p) = p$, (рис. 3).

Примеры р-квантили:

$x_{0,5}$ — медиана — характеристика среднего значения случайной величины;

Рис. 3. Графическая иллюстрация квантили x_p

$x_{0,98}$ — максимальное, с вероятностью 0,98, значение случайной величины, т.к. $P\{\xi < x_{0,98}\} = 0,98$;

$x_{0,02}$ — минимальное, значение случайной величины, т.к. $P\{\xi \geq x_{0,02}\} = 1 - P\{\xi < x_{0,02}\} = 1 - 0.02 = 0,98$;

$x_{3/4}$ и $x_{1/4}$ — верхняя и нижняя квартили; их разность $(x_{0,75} - x_{0,25})$ — межквартильная широта — служит характеристикой разброса.

Оценка p-квантилей. Неизвестные *p*-квантили легко оцениваются по выборке. Действительно, пусть $x_1, x_2 \ldots x_n$ —выборка, результаты *n* независимых наблюдений над случайной величиной ξ с функцией распределения $F(x)$. Упорядочив их по возрастанию, получаем вариационный ряд

$$x_{(1)} \leq x_{(2)} \leq \ \ldots \leq x_{(n)}.$$

Чтобы подчеркнуть случайность ряда, запишем его греческими символами

$$\xi_{(1)} \leq \xi_{(2)} \leq \ \ldots \leq \xi_{(n)}.$$

Член вариационного ряда $\xi_{(i)}$ с номером i (заметим, что это случайная величина) называется *i-й порядковой статистикой*. По вариационному ряду построим функцию

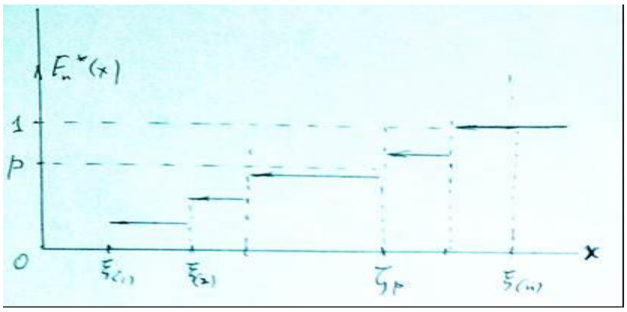
$$F_n^*(x) \equiv F_n^*(x; \xi_{(1)} \leq \xi_{(2)} \leq \ldots \leq \xi_{(n)})$$

эмпирического распределения, и, согласно общему принципу о том, что выборочные характеристики являются состоятельными оценками характеристик распределения генеральной совокупности, рассмотрим в качестве оценки для p-квантили x_p выборочную квантиль ζ_p, т.е. корень уравнения

$$F_n^*(\zeta_p) = p. \qquad (8)$$

Поскольку $F_n^*(x)$ — функция кусочно-постоянная, то корнем является одна из порядковых статистик $\zeta_p = \xi_{([np]+1)}$, (9)

с номером $\left[\frac{p}{1/n}\right]+1 = [np]+1$, т.е. целая часть числа np плюс 1(рис. 4).



Нетрудно показать, что ζ_p *является состоятельной оценкой для* x_p:

$$\zeta_p \xrightarrow[n\to\infty]{P} x_p$$

Кроме того, известна **теорема Крамера**, которая гласит: *для непрерывных распределений с плотностью* *q(x) оценка* ζ_p *асимптотически нормальна с параметрами*:

$$M\zeta_p = x_p, \quad D\zeta_p = \frac{1}{n}\frac{p(1-p)}{q^2(x_p)}. \qquad (10)$$

Рис. 4. Графическая иллюстрация выборочной квантили

Метод оценки параметров основан на оценках ζ_p при разных p. Как в методе моментов? параметры выражаем через моменты, а затем моменты заменяем выборочными моментами. Аналогично в методе порядковых статистик: параметры выражаем через квантили, а затем квантили заменяем выборочными квантилями, т.е. порядковыми статистиками.

Пусть $\xi_1, \xi_2 \ldots \xi_n$ — выборка с функцией распределения $F(x;a)$, зависящей от параметра a, значение которого требуется оценить. Выберем p так, чтобы квантиль x_p зависела от параметра:

$$x_p = f(a).$$

Выразим параметр a через квантиль x_p:

$$a = g(x_p),$$

и вместо x_p подставим выборочную квантиль $\zeta_p = \xi_{([np]+1)}$, в результате чего получим *состоятельную оценку*

$$\hat{a} = g(\xi_{([np]+1)}).$$

Таким же образом можно построить оценки и для неодномерного параметра.

Основное и очень важное преимущество оценок, основанных на порядковых статистиках, — *их устойчивость к засорению наблюдений и к изменениям закона распределения*.

Примеры оценок параметров нормального распределения. Пусть $\xi_1, \xi_2 \ldots \xi_n$ — выборка из нормальной совокупности $N(m, \sigma^2)$.

1) Оценка среднего m. Известно или нет значение σ — безразлично. В силу симметрии нормального распределения параметр *m* является медианой, т.е. квантилью уровня ½,

$$m = x_{1/2}$$

и потому может быть оценен выборочной медианой:

$$\hat{m} = \zeta_{1/2} = \xi_{([n/2]+1)}.$$

Можно сравнить по точности эту оценку с эффективной оценкой

$$\hat{m} = \sum_{i=1}^{n} \xi_i / n$$

для которой дисперсия $$\mathrm{D}\hat{m}^* = \sigma^2 / n.$$

Согласно (10), теореме Крамера, $$\mathrm{D}\hat{m} \approx \frac{1}{n} \cdot \frac{0{,}5(1-0{,}5)}{(2\pi\sigma^2)^{-1}} = \frac{\sigma^2}{n} \cdot \frac{\pi}{2},$$

т.е. очень простая и устойчивая к засорению оценка $\hat{m}$ уступает по точности оценке $\hat{m}^*$ в $\sqrt{\pi/2} \approx 1{,}25$ раза, т.е. 25 %.

2) Оценка стандартного уклонения σ.

Легко проверить, что верхняя и нижняя квартили равны соответственно

$$x_{3/4} = m + 0{,}675\sigma \quad \text{и} \quad x_{1/4} = m - 0{,}675\sigma,$$

т.к. $P\{\xi < m + 0.675\sigma\} = 3/4,\ P\{\xi < m - 0.675\sigma\} = 1/4$

И потому

$$\sigma = (x_{3/4} - x_{1/4}) / 1{,}35,$$

и потому оценивать σ можно следующим образом:

$$\hat{\sigma} = \frac{\zeta_{3/4} - \zeta_{1/4}}{1{,}35} = \frac{\xi_{([3n/4]+1)} - \xi_{([n/4]+1)}}{1{,}35}.$$

3) Оценка стандартного уклонения σ *по размаху*.

Пусть $\xi_{(1)}$ и $\xi_{(n)}$ — минимальный и максимальный член выборки, разность которых называется размахом *w*:

$$w = \xi_{(n)} - \xi_{(1)}.$$

Ясно, что

$$\mathrm{M}w = c(n)\sigma,$$

и потому оценкой для σ может служить

$$\hat{\sigma} = w/c(n) = k(n)w,$$

где $k(n)$ берем из статистических таблиц [4]. Ниже приведены значения коэффициента $k(n)$ и коэффициента эффективности

$$eff = \frac{\sigma_0^2}{D\hat{\sigma}},$$ где σ_0^2 — нижняя граница Рао-Крамера,

а также потеря точности:

$$(1 - \sqrt{eff}) \cdot 100,$$

измеряемая в процентах, по сравнению с нижней границей Рао-Крамера.

Табл. 1. Значение коэффициентов *k* и *n*

n	2	5	10
k(n)	0,866	0,430	0,325
eff	1,000	0,955	0,855
потеря точности, (1 – $\sqrt{eff}$)100, %	0	2,5	7

Для устойчивости оценки к засорению используют подразмахи w_m порядка *m,* где *m* = 1, 2, 3…:

$$w_m = \xi_{(n-m+1)} - \xi_{(m)},$$

так что оценка имеет вид:

$$\hat{\sigma} = k_m(n)\, w_m.$$

Значение коэффициента $k_m(n)$ берется из таблиц.

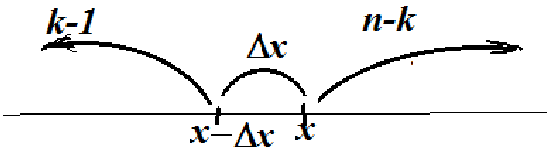
4) Распределение порядковых статистик. При анализе оценок, получаемых рассматриваемым методом, необходимо знать распределения порядковых статистик. Если распределение одного наблюдения ξ непрерывно с плотностью $p(x) = F'(x)$, то плотность распределения для *k*-й порядковой статистики $\xi_{(k)}$ выражается следующей формулой:

$$p_{\xi_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} F^{k-1}(x)[1-F(x)]^{n-k}\, p(x),$$

которая получается вычислением вероятности события

$$p_{\xi_{(k)}}(x)\Delta x = P\{\xi_{(k)} \in (x-\Delta x, x)\},$$

по полиномиальной схеме. Событие означает, что при *n*-кратном испытании случайной величины ξ



событие $\{\xi < x-\Delta x\}$, вероятность которого $F(x-\Delta x)$, появится (*k*-1) раз: множитель $F^{k-1}(x)$,

событие $\{\xi \geq x\}$, вероятность которого $(1-F(x))$, появится (*n*-*k*) раз: множитель $[1-F(x)]^{n-k}$

и событие $\{\xi \in (x-\Delta x, x)\}$, вероятность которого $p(x)\Delta x$, появится 1 раз: множитель $p(x)$.

ЧАСТЬ 2. ИНТЕРВАЛЬНОЕ ОЦЕНИВАНИЕ

§ 6. Доверительные границы и интервалы

Результатом применения оценки $\hat{a}(x_1, x_2 ... x_n)$ является одно числовое значение, которое не дает представления о точности, т.е. о том, насколько близко полученное значение к истинному значению параметра. Интуитивно ясно, что такое представление может дать, например, дисперсия оценки, так что истинное значение неизвестного должно находиться где-то в пределах

$$\hat{a} \pm (2 \div 4)\sqrt{D\hat{a}}.$$

Внесем уточнения того, что мы хотим.

6.1. Определения

Пусть $\xi \equiv (\xi_1, \xi_2 ... \xi_n)$ —выборка, т.е. n независимых наблюдений над случайной величиной с функцией распределения $F(x;a)$, зависящей от параметра a, значение которого неизвестно.

Определение 1. Интервал $I(\xi) = (a_1(\xi), \ a_2(\xi))$ со случайными концами (случайный интервал), определяемый двумя функциями наблюдений, называется *доверительным интервалом* для параметра a с уровнем доверия $P_Д$ (обычно близким к 1), если

$$\min_a \mathbf{P}\{I(\xi) \ni a\} \equiv \min_a \mathbf{P}\{a_1(\xi) < a < a_2(\xi)\} = P_Д, \tag{1}$$

т.е. если при любом значении параметра a *минимальная вероятность* (зависящая от a) *накрыть случайным интервалом* $I(\xi)$ *истинное значение* a *велика* (не менее $P_Д$), и равна $P_Д$.

Определение 2. Функция наблюдений $\hat{a}_н(\xi_1, \xi_2 ... \xi_n)$ (случайная величина), называется *нижней доверительной границей* для параметра a с уровнем доверия $P_Д$ (близким к 1), если

$$\min_a \mathbf{P}\{\hat{a}_н(\xi_1, \xi_2 ... \xi_n) < a\} = P_Д. \tag{2}$$

Т.е. если при любом значении параметра a *минимальная* вероятность события $\{\hat{a}_н(\xi) < a\}$ велика (не меньше $P_Д$), и равна $P_Д$.

Определение 3. Функция наблюдений $\hat{a}_в(\xi_1, \xi_2 ... \xi_n)$ (случайная величина) называется *верхней доверительной границей* для параметра a с уровнем доверия $P_Д$, если

$$\min_a \mathbf{P}\{\hat{a}_в(\xi_1, \xi_2 ... \xi_n) > a\} = P_Д. \tag{3}$$

Т.е. если при любом значении параметра a *минимальная* вероятность события $\{\hat{a}_в(\xi) > a\}$ велика (не меньше $P_Д$), и равна $P_Д$.

Вероятность $P_Д$ называют также *доверительной вероятностью*.

6.2. Пример. Доверительный интервал для среднего нормальной совокупности при известной дисперсии

Пусть $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ — выборка из нормальной $N(a, \sigma^2)$ совокупности. Достаточной оценкой для a является

$$\hat{a} = \hat{a}(\xi_1, \xi_2 \ldots \xi_n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i \equiv \bar{\xi},$$

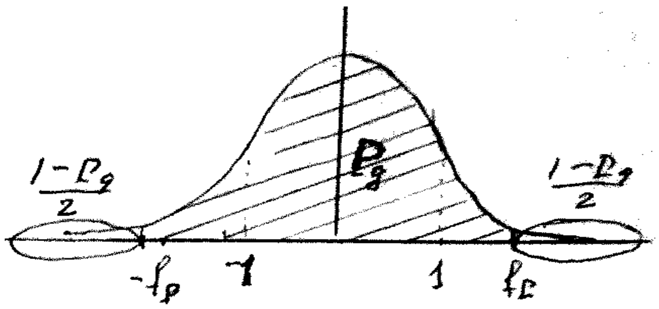
распределенная по нормальному закону $N(a, \frac{\sigma^2}{n})$. Пронормируем её, образовав случайную величину

$$\varphi(\bar{\xi}; a) = \frac{\bar{\xi} - a}{\sigma/\sqrt{n}}, \qquad (4)$$

которая распределена нормально $N(0,1)$ ***при любом значении a.***

По заданному уровню доверия $P_Д$ (рис. 5) определим для φ симметричный интервал $(-f_p, f_p)$ так, чтобы он содержал в себе вероятность $P_Д$, т.е.

$$\mathbf{P}\{-f_p < \varphi < f_p\} = P_Д. \qquad (5)$$



Ясно, что f_p есть квантиль порядка $(1 + P_Д) / 2$ стандартного нормального распределения $N(0,1)$. Заметим, что φ зависит от a, и (5) верно при любом значении a.

Рис. 5. Выбор интервала при нормальном распределении

Подставим в (5) выражение для φ из (4) и разрешим неравенство под знаком вероятности в (5) относительно a. Получим соотношение

$$\mathbf{P}\left\{\bar{\xi} - f_p\frac{\sigma}{\sqrt{n}} < a < \bar{\xi} + f_p\frac{\sigma}{\sqrt{n}}\right\} = P_Д \qquad (6)$$

верное по-прежнему при любом значении a. Под знаком вероятности слева и справа имеем две функции наблюдений

$$a_1(\xi_1, \xi_2 \ldots \xi_n) \equiv \bar{\xi} - f_p\frac{\sigma}{\sqrt{n}} \quad \text{и} \quad a_2(\xi_1, \xi_2 \ldots \xi_n) \equiv \bar{\xi} + f_p\frac{\sigma}{\sqrt{n}}, \qquad (7)$$

определяющие случайный интервал

$$I(\xi_1, \xi_2 \ldots \xi_n) = (a_1(\xi), a_2(\xi)), \qquad (8)$$

который в силу (6) накрывает неизвестное значение параметра a с большой вероятностью, равной $P_Д$, при любом значении параметра a, и потому, по определению доверительного интервала, он является доверительным для a с уровнем доверия $P_Д$.

6.3. Способ построения доверительных границ и интервалов

Для построения доверительного интервала (или границы) необходимо знать ***закон распределения статистики*** $\zeta = \zeta(\xi_1, \xi_2 ... \xi_n)$***, по которой оценивается неизвестный параметр*** (такой статистикой могут быть сама оценка $\hat{a}(\xi_1, \xi_2 ... \xi_n)$, статистика, от которой зависит оценка $\hat{a}$, достаточная статистика или статистика, близкая к достаточной).

Один из способов построения состоит в следующем (проследим логику рассмотренного примера).

1) Построим случайную величину $\varphi = \varphi(\zeta, a)$, зависящую от статистики ζ и неизвестного параметра a таким образом, что:

— закон распределения для φ известен и не зависит от a;

— $\varphi(\zeta, a)$ непрерывна и монотонна по a.

Такая случайная величина $\varphi(\zeta, a)$ называется *центральной статистикой*.

2) Выберем интервал (f_1, f_2) для φ так, чтобы попадание в него случайной величины φ было практически достоверным (с вероятностью $P_Д$):

$$P\{f_1 < \varphi(\zeta, a) < f_2\} = P_Д, \qquad (9)$$

для чего достаточно в качестве f_1 и f_2 взять квантили распределения для φ уровня $(1-P_Д)/2$ и $(1+P_Д)/2$ соответственно (рис. 6).

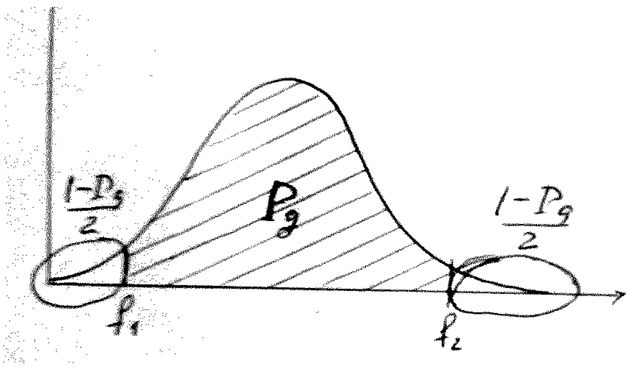


Рис. 6. Выбор интервалов

3) Перейдем в (9) к другой записи случайного события, разрешив неравенства относительно параметра a.

Предполагая *монотонное возрастание* φ по a получим:

$$P\{g(\zeta, f_1) < a < g(\zeta, f_2)\} = P_Д.$$

Это соотношение верно при любом значении параметра a (поскольку это так для (9)), и потому, согласно определению, интервал со случайными концами $(g(\zeta, f_1), g(\zeta, f_2))$ является доверительным для a с уровнем доверия $P_Д$.

Если имеем монотонное *убывание* φ по a, интервалом будет

$$(g(\zeta, f_2), g(\zeta, f_1)).$$

Замечания.

1. Сделанное выше предположение о монотонности $\varphi(\zeta, a)$ по a, не является существенным. В случае, если монотонности нет,

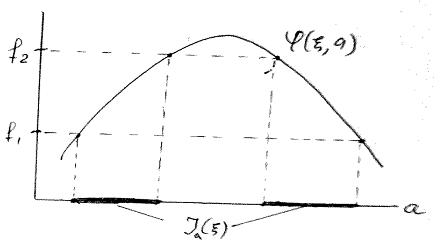


Рис. 7. Доверительное множество при отсутствии монотонности

результатом разрешения неравенств в (9) под знаком вероятности относительно параметра a является не интервал, а более сложное *доверительное множество*, например, два интервала (рис. 7).

2. Если требуется построить *односторонною доверительную границу* (верхнюю или нижнюю), нужно использовать только одно из неравенств под знаком вероятности в (9). Используем

$$\text{или}\quad \mathbf{P}\{\varphi(\zeta, a) > f_1\} = P_Д, \quad f_1 = Q(1 - P_Д),$$

$$\text{или}\quad \mathbf{P}\{\varphi(\zeta, a) < f_2\} = P_Д, \quad f_2 = Q(P_Д),$$

согласовав знак неравенства с характером монотонности φ по a (возрастание или убывание) и с характером границы (верхняя или нижняя). В неравенствах обозначено $Q(P)$ — квантиль уровня P. После разрешения неравенства под знаком **P** получим односторонние доверительные границы для a.

3. О числовых значениях интервала. После применения формул для интервала, например, (8), к конкретным наблюдениям мы получаем конкретный интервал, например, для $n = 9$, при $\sigma = 1$, $P_Д = 0{,}95$ (соответствующее значение $f_P = 1{,}96$) и $\sum_{i=1}^{n} x_i \Big/ n = 3{,}87$ получаем по (8)

$$a_1 = 3{,}22,\ a_2 = 4{,}52.$$

Смысл замечания состоит в том, что нельзя писать аналогично соотношению (6)

$$\cancel{\mathbf{P}\{3.22 < a < 4.52\} = P_Д},$$

поскольку под знаком вероятности нет случайного события. Параметр a не является случайной величиной, и интервал (3,22, 4,52) тоже не случайный, в отличие от формулы (6), где интервал является случайным. Коэффициент доверия $P_Д = 0{,}95$ — это *характеристика способа* определения интервала, но не конкретного полученного интервала.

4. О точности интервального оценивания (о ширине интервала (6)).

$$\mathbf{P}\left\{\bar{\xi} - f_p \frac{\sigma}{\sqrt{n}} < a < \bar{\xi} + f_p \frac{\sigma}{\sqrt{n}}\right\} = P_Д$$

(6)

а) Ошибка оценивания δ, в соответствии с (6),

$$\delta = \left|\bar{\xi} - a\right| = \left|\hat{a} - a\right| < f_p \frac{\sigma}{\sqrt{n}}$$

не превосходит значения $f_p \sigma/\sqrt{n}$ с вероятностью $P_Д < 1$. При массовых вычислениях подобного рода точность не хуже $f_p \sigma/\sqrt{n}$ в числе случаев, доля которых $P_Д$. Уровень доверия $P_Д$ означает, что правило определения интервала дает верный результат с вероятностью $P_Д$, которая обычно близка к единице, однако, единице не равна.

б) Хотелось бы иметь высокую точность с высокой вероятностью. Однако, при увеличении $P_Д$ значение f_p тоже увеличивается, так что с ***высокой вероятностью можно гарантировать относительно низкую точность*** — чем надежнее гарантия, тем меньше она гарантирует.

в) Связь точности с σ и *n*: чем меньше σ, тем выше точность. Увеличение *n* также улучшает точность, которая изменяется как $1/\sqrt{n}$. Чтобы повысить точность в 3 раза, нужно число наблюдений увеличить в 9 раз.

6.4. Интервалы для параметров нормального распределения

А. Распределение хи-квадрат с k степенями свободы. Для рассмотрения типичных практических примеров потребуются сведения о некоторых распределениях. Многие задачи статистики связаны с распределением хи-квадрат ($\chi^2(k)$).

Пусть α_1, $\alpha_2 \ldots \alpha_k$ — независимые случайные величины, распределенные по стандартному нормальному закону *N*(0,1). Рассмотрим сумму их квадратов и обозначим соответствующую случайную величину через χ_k^2:

$$\chi_k^2 = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_k^2. \tag{10}$$

Распределение этой случайной величины называют *распределением хи-квадрат с k степенями свободы*. Нетрудно показать (см., например, [2], §24), что плотность этого распределения выражается следующей формулой:

$$h_k(x) = C_k x^{k/2-1} e^{-x/2}, \ x > 0, \tag{11}$$

где $C_k = \left(2^{k/2}\Gamma(k/2)\right)^{-1}$ — нормирующий множитель, $\Gamma(\lambda) = \int\limits_0^\infty t^{\lambda-1}e^{-t}dt$ — гамма-функция. На рис. 8 показаны графики при различных значениях k.

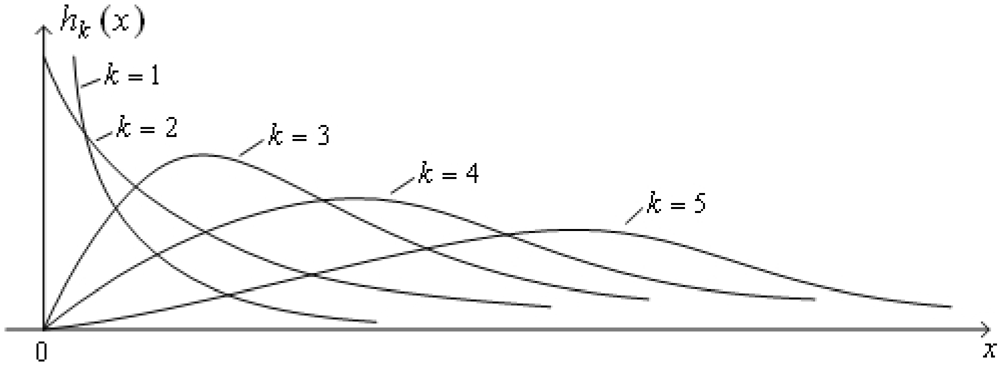


Рис. 8. Семейство плотностей распределения χ^2

Заметим, что при k = 2 получаем показательное распределение:

$$h_2(x) = 0{,}5e^{-x/2}.$$

Из соотношения (10) получаем первые два момента:

$$\mathrm{M}\chi_k^2 = k, \qquad \mathrm{D}\chi_k^2 = 2k, \text{ (проверить!)}$$

откуда ясно, что с увеличением числа k степеней свободы распределение $\chi^2(k)$ смещается вправо и расплывается, а также, что оно асимптотически нормально (в силу центральной предельной теоремы):

$$\chi^2(k) \sim N(k, 2k) \text{ при } k\to\infty;$$

при k > 30 можно пользоваться таблицами нормального распределения.

Далее отметим весьма полезные сведения.

Замечание. Это распределение χ^2 является частным случаем *гамма-распределения*, для которого плотность выражается формулой

$$p(x;\lambda,a) = C(\lambda,a)\,x^{\lambda-1}e^{-ax},\ x>0,\ \lambda>0,\ a>0,$$

где $C(\lambda,a) = a^\lambda/\Gamma(\lambda)$ - нормирующий множитель; λ – параметр формы, a – параметр масштаба, $\Gamma(\lambda) = \int\limits_0^\infty t^{\lambda-1}e^{-t}dt$ - известная гамма-функция.

Первые два момента m_1 и σ^2 равны соответственно

$$m_1 = \lambda/a,\ \sigma^2 = \lambda/a^2.$$

Характеристическая функция *f(t)* этого распределения выражается формулой:

$$f(t) = Me^{it\xi} = \int\limits_0^\infty e^{itx}p(x;\lambda,a)dx = \frac{a^\lambda}{\Gamma(\lambda)}\int\limits_0^\infty e^{itx}x^{\lambda-1}e^{-ax}dx = \left(1-i\frac{t}{a}\right)^{-\lambda}. \qquad (12)$$

Если λ — целое число, то распределение называется *распределением Эрланга*, которому подчиняется сумма λ независимых случайных величин, показательно распределенных с плотностью $ae^{-ax}, x > 0$.

Справедливость формулы (11) можно легко показать, определив характеристические функции для α_1^2 и затем для $\chi_k^2 = \alpha_1^2 + \alpha_2^2 + ... + \alpha_k^2$. Характеристическая функция для случайной величины χ_k^2 оказывается равной

$$(1-2it)^{-k/2},$$

откуда следует, что соответствующее распределение является гамма-распределением с параметрами $\lambda = k/2$, $a = 1/2$.