МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Лекция 6

Повтор.

Определение. 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

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

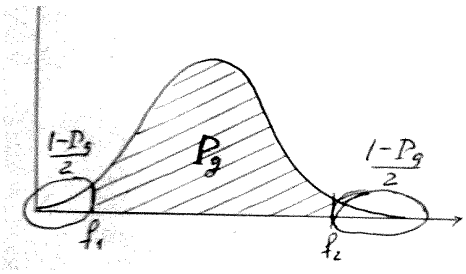
т.е. если при любом значении параметра a *вероятность* (зависящая от a) *накрыть случайным интервалом* $I(\xi)$ *истинное значение* a велика, не менее заданной величины $P_Д$.

Один из способов был виден из примера: оценка среднего нормальной совокупности. Пусть $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ $N(a, \sigma^2)$ совокупности, a=?, σ известно.

$\hat{a} = \hat{a}(\xi_1, \xi_2 \ldots \xi_n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i \equiv \overline{\xi}$- оценивающая статистика, з.р. известен $N(a, \frac{\sigma^2}{n})$ (3)

$\overline{\xi} \equiv \zeta$ - оценивающая статистика (можно взять $\sum_{i=1}^{n}\xi_i$ -тоже оценивающая ст-ка)

1.Конструируем с.в. φ введением параметра a так, чтобы з.р. был известен: $\varphi(\zeta; a) = \frac{\zeta - a}{\sigma/\sqrt{n}}$, (4)



2. По заданному уровню доверия $P_Д$ определим для φ интервал (f_1, f_2) так, чтобы он содержал в себе вероятность $P_Д$, т.е.

$$\mathbf{P}\{f_1 < \varphi < f_2\} = P_Д$$

$$\frac{\zeta - a}{\sigma/\sqrt{n}} \quad . \ \forall a \qquad (5)$$

3). Разрешаем неравенства под знаком вер-ти :

$$\mathbf{P}\{g_1(\zeta) < a < g_2(\zeta)\} = P_Д \qquad (6)$$

Теперь под знаком вероятности стоит событие, состоящее в том, что случайный интервал накроет неизвестное значение параметра с заданной большой вероятностью $P_Д$ при любом значении параметра, т.е. $(g_1(\zeta), g_2(\zeta))$- доверит. инт-л с уровнем доверия $P_Д$.

Конец повтора

Доверительные границы -2
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^2_k:

$$\chi^2_k = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_k^2 . \qquad (10)$$

Распределение этой случайной величины называют

распределением хи-квадрат с k степенями свободы.

Нетрудно показать (см., например, [2], Гнеденко, Курс теории вероятностей, §24), что плотность этого распределения выражается следующей формулой:

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

где $C_k = \left(2^{k/2}\Gamma(k/2)\right)^{-1}$ — нормирующий множитель, $\Gamma(\lambda) = \int_0^\infty t^{\lambda-1} e^{-t}\,dt$ — знаменитая гамма-функция; напомним, что $\Gamma(\lambda) = (\lambda-1)\Gamma(\lambda\text{-}1)$, и при целом λ, $\Gamma(\lambda) = (\lambda-1)!$.

На рис. 8 показаны графики при различных значениях k.

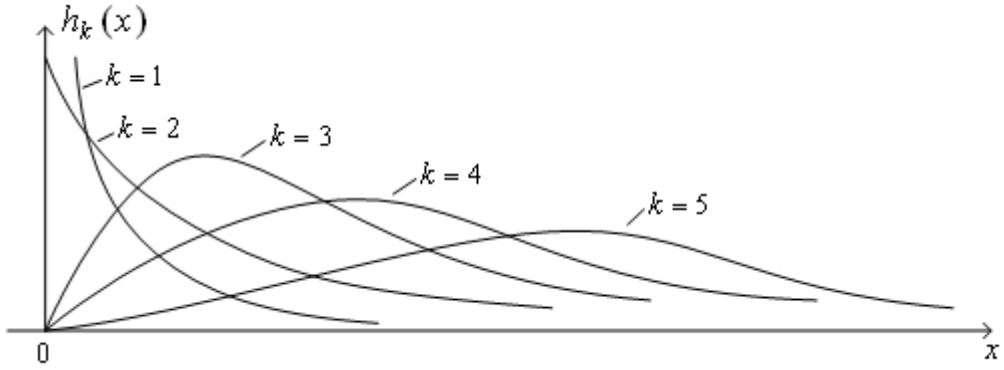


Рис. 8. Семейство плотностей распределения χ^2

Заметим, что при $k = 2$ получаем показательное распределение:

$$h_2(x) = 0{,}5e^{-x/2} \quad \sim \quad E(1/2).$$

Из соотношения (10) получаем первые два момента:

$$\mathrm{M}\chi^2_k = k, \qquad \mathrm{D}\chi^2_k = 2k,$$

Проверяем: $\mathrm{D}\,\chi^2_k = \sum_{i=1}^{k} D\alpha_i^2 = kD\alpha_i^2 = k\left[M\alpha_i^4 - \left(M\alpha_i^2\right)^2\right] = k\left(3-1^2\right) = 2k$

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 полезные сведения.

Замечание о связи с гамма-распределением. Распределение χ^2- хи-квадрат является частным случаем *гамма-распределения*, для которого плотность выражается формулой

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

где $C(\lambda, a) = a^{\lambda}/\Gamma(\lambda)$ - нормирующий множитель; λ – параметр формы, a – параметр масштаба, $\Gamma(\lambda) = \int_0^{\infty} t^{\lambda-1}e^{-t}dt$ - 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:

$$f(t) = Me^{it\xi} = \int_0^{\infty} e^{itx} p(x;\lambda,a)dx = \frac{a^{\lambda}}{\Gamma(\lambda)}\int_0^{\infty} e^{itx}x^{\lambda-1}e^{-ax}dx = \frac{1}{\Gamma(\lambda)}\int_0^{\infty} e^{-\left(1-i\frac{t}{a}\right)ax}(ax)^{\lambda-1}d(ax) =$$

новая переменная интегрирования: $z = \left(1-i\frac{t}{a}\right)ax$

$$= \frac{\left(1-i\frac{t}{a}\right)^{-\lambda}}{\Gamma(\lambda)}\int_0^{\infty} e^{-z}z^{\lambda-1}dz = \left(1-i\frac{t}{a}\right)^{-\lambda}. \qquad (12)$$

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 $\lambda = k/2$, $a = 1/2$.

Б. *Совместное распределение выборочных среднего и дисперсии нормальной совокупности*. ***(важный вопрос! обобщнеие в регрессионном анализе)***

Теорема. Пусть $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ — выборка $N(m, \sigma^2)$, оценки параметров:

$\bar{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i$ —$s^2=\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\bar{\xi})^2$ выборочные среднее и дисперсия.

Утверждения:

1) эти статистики $\bar{\xi}$ и s^2 независимы;

2) с. в. $\sqrt{n}(\bar{\xi}-m)/\sigma \sim N(0, 1)$- стандартный нормальный закон,

3) $ns^2/\sigma^2 \sim \chi^2(n\text{-}1)$ — хи-квадрат с числом степеней свободы (n-1).

Доказательство. Перейдем нормировкой к новым случайным величинам

$\eta_i = (\xi_i - m)/\sigma$, $i = 1, 2 \ldots n$, которые образуют выборку $\eta = (\eta_1, \eta_2 \ldots \eta_n)$ из совокупности, распределенной по $N(0, 1)$. Тогда

$$\bar{\eta}=\frac{1}{n}\sum_{i=1}^{n}\eta_i = (\bar{\xi}-m)/\sigma,$$

Далее:

$$\frac{ns^2}{\sigma^2}=\sum_{i=1}^{n}\left(\frac{\xi_i-\bar{\xi}}{\sigma}\right)^2=\sum_{i=1}^{n}\left(\frac{\xi_i-m}{\sigma}-\frac{\bar{\xi}-m}{\sigma}\right)^2=n\left[\frac{1}{n}\sum_{i=1}^{n}(\eta_i-\bar{\eta})^2\right]=n\left(\frac{1}{n}\sum_{i=1}^{n}\eta_i^2-\bar{\eta}^2\right)=\sum_{i=1}^{n}\eta_i^2-n\bar{\eta}^2 \quad (13)$$

здесь предпоследняя сумма есть умноженная на n дисперсия выборочного распределения.

Преобразуем вектор η с помощью ортогонального преобразования с матрицей C:

$$\zeta = C\eta,$$

где первая строка матрицы C состоит из одинаковых элементов, равных $1/\sqrt{n}$. Дисперсионная матрица ζ, с учетом того, что

$$M(\eta\eta^T) = I \text{ и } C^T = C^{-1},$$

равна

$$D\zeta = MC\eta(C\eta)^T = CM(\eta\eta^T)C^T = I,$$

где I — единичная матрица, и потому $\zeta_1, \zeta_2 \ldots \zeta_n$ — независимые случайные величины, распределенные по $N(0, 1)$.

Если учесть, что ортогональное преобразование не меняет расстояния, т. е. $\sum_{i=1}^{n}\eta_i^2=\sum_{i=1}^{n}\zeta_i^2$,

а для первого элемента справедливо соотношение

$$\zeta_1=\sum_{i=1}^{n}\frac{1}{\sqrt{n}}\eta_i = \sqrt{n}\bar{\eta},$$

то выражение (13) примет вид

$$ns^2/\sigma^2 = \sum_{i=1}^{n}\eta_i^2-n\bar{\eta}^2=\sum_{i=1}^{n}\zeta_i^2-\zeta_1^2 = \zeta_2^2+\zeta_3^2+\ldots+\zeta_n^2.$$

Последняя сумма ns^2/σ^2 распределена по закону хи-квадрат с (n–1) степенями свободы и не зависит от

$$\zeta_1 = \sqrt{n}\,\eta = \sqrt{n}\,\frac{\bar{\xi} - m}{\sigma}$$, т.е. от $\bar{\xi}$.

Именно это утверждает данная теорема.

В. Доверительный интервал для дисперсии нормальной совокупности. Пусть $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ — выборка из совокупности, распределенной по нормальному закону $N(m, \sigma^2)$. Задан коэффициент доверия $P_Д$.

Параметр m может быть известен или не известен, поэтому рассматриваем два случая одновременно. В качестве несмещенных оценок для σ^2 используем статистики:

если m известно, то

$$s^2 = \sum_{i=1}^{n} (\xi_i - m)^2 \Big/ n$$, и тогда $$\frac{ns^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\xi_i - m}{\sigma} \right)^2 \sim \chi^2_n,$$

иначе

$$s^2 = \sum_{i=1}^{n} \left(\xi_i - \bar{\xi} \right)^2 \Big/ (n-1)$$, и тогда $$\frac{(n-1)s^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{\xi_i - \bar{\xi}}{\sigma} \right)^2 \sim \chi^2_{n-1}$$ по теореме.

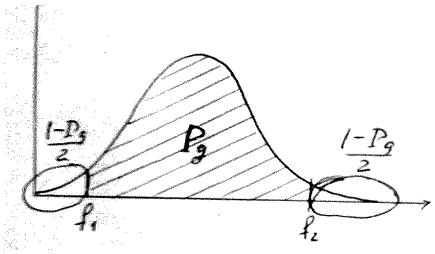
Рассмотрим случайную величину

$$\varphi(\xi, m, \sigma) = ks^2 / \sigma^2$$, где $$k = \begin{cases} n, \text{ если } m \text{ известно,} \\ (n-1), \text{ если } m \text{ неизвестно.} \end{cases}$$

Очевидно, что в обоих случаях случайная величина φ подчиняется закону распределения хи-квадрат с k степенями свободы.

Определим интервал (f_1, f_2) так, чтобы

$$\mathbf{P}\{f_1 < \varphi < f_2\} = P_Д.$$



В качестве f_1 и f_2 возьмем квантили уровней соответственно $(1 - P_Д) / 2$ и $(1 + P_Д) / 2$ распределения хи-квадрат с k степенями свободы:

$$\mathbf{P}\{\chi^2_k < f_1\} = (1\text{-}P_Д)/2, \quad \mathbf{P}\{\chi^2_k < f_2\} = (1+P_Д)/2.$$

Разрешая под знаком вероятности два неравенства относительно σ

$$f_1 < \varphi = ks^2 / \sigma^2 < f_2,$$

получим соотношение

$$\mathbf{P}\left\{ s\sqrt{k/f_2} < \sigma < s\sqrt{k/f_1} \right\} = P_Д$$

верное при любых значениях m и σ, откуда следует, что интервал $\left(s\sqrt{k/f_2},\ s\sqrt{k/f_1} \right)$ является доверительным для σ с доверительной вероятностью $P_Д$.

Пример. Пусть среднее m неизвестно, $n = 2$, $P_Д = 0{,}95$. Тогда

$$s = |x_1 - x_2| / \sqrt{2} = \frac{\Delta x}{\sqrt{2}},$$

и доверительный интервал весьма широк — (0,5*s*, 30*s*).
При *n* = 10 : (0,7*s*, 1,8*s*),
при *n* = 20 — (0,87*s*, 1,17*s*).
Вывод: если оцениваете с.к.о., не верьте точечной оценке, обязательно считайте доверит. интервал.

***Г. Распределение Стьюдента*.** Многие задачи статистики приводят к рассмотрению следующей случайной величины.

Пусть с.в.: α, $N(0,1)$,

и с.в. χ^2_k, ~ хи-квадрат с *k* степенями свободы.

Образуем новую случайную величину T_k следующим образом:

$$T_k = \frac{\alpha}{\sqrt{\chi^2_k / k}}. \qquad (14)$$

Распределение этой случайной величины называется ***распределением Стьюдента*** (псевдоним английского статистика В. Госсета) с *k* степенями свободы и обозначается $s(k)$. Плотность $s_k(x)$ распределения выражается формулой:

$$s_k(x) = C_k \frac{1}{\left(1 + x^2/k\right)^{(k+1)/2}}, \; -\infty < x < \infty,$$

где $C_k = \frac{1}{\sqrt{\pi k}} \frac{\Gamma((k+1)/2)}{\Gamma(n/2)}$ — нормирующий множитель.

При *k* = 1 распределение Коши с плотностью $\frac{1}{\pi(1+x^2)}$.

При увеличении *k* знаменатель в (14) сходится к 1, поскольку математическое ожидание

$\mathrm{M}\chi^2_k / k = 1$, а дисперсия $\mathrm{D}\left[\chi^2_k / k\right] \to 0$,

и потому распределение $S(k)$ сходится к стандартному нормальному.

При *k* > 30 для вероятностей p > 0,01 можно нормальным распределением.

$$T_k \sim N(0, 1);$$

Д. Доверительный интервал для среднего нормальной совокупности при неизвестной дисперсии.

Пусть $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ — выборка из совокупности, распределенной по нормальному закону $N(m, \sigma^2)$. Построим доверительный интервал с коэффициентом доверия $P_Д$. Параметр σ неизвестен, но именно он определяет точность оценки, т.е. ширину интервала, поэтому его тоже нужно оценить.

Пусть $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ — оценка м.о. $\sim N(m, \frac{1}{n}\sigma^2.)$

$\alpha = \dfrac{\overline{\xi} - m}{\sigma / \sqrt{n}} \sim N(0, 1)$ - стандартный нормальный закон,

$s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2$ — оценка дисперсии,

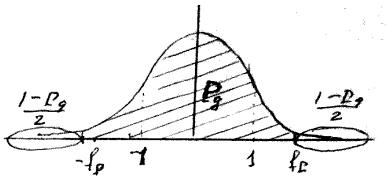
$\dfrac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1}$ - распределение хи-квадрат с (n–1) степенями свободы,

и α и s^2 независимы.

Построим статистику $T_{n\text{-}1}$ делением

$$T_{n\text{-}1} = \left(\frac{\overline{\xi} - m}{\sigma / \sqrt{n}}\right) \Bigg/ \sqrt{\frac{(n-1)s^2}{\sigma^2} \Big/ (n-1)} = \frac{\overline{\xi} - m}{s / \sqrt{n}} \forall m, \sigma^2 = \varphi$$

Неизвестное значение σ сократилось. В силу определения с.в. с законом Стьюдента и теоремы о независимости выборочных среднего и дисперсии нормальной совокупности, эта статистика подчиняется закону Стьюдента с (n–1) степенями свободы. По заданному коэффициенту доверия $P_Д$ определяем симметричный интервал (- t_P, t_P) такой, что



$$\mathbf{P}\{- t_P < T_{n\text{-}1} < t_P \} = P_Д \ \forall m, \sigma^2$$

Очевидно, что t_P есть квантиль уровня (1+$P_Д$) / 2.

Разрешая под знаком вероятности два неравенства относительно параметра m

$$- t_P < \frac{\overline{\xi} - m}{s / \sqrt{n}} < t_P,$$

получаем:

$$\mathbf{P}\{ \overline{\xi} - t_P \frac{s}{\sqrt{n}} < m < \overline{\xi} + t_P \frac{s}{\sqrt{n}} \} = P_Д \ \forall m, \sigma^2$$

Последнее соотношение верно при любых значениях параметров m и σ, и потому случайный интервал $\{ \overline{\xi} - t_P \frac{s}{\sqrt{n}}, \ \overline{\xi} + t_P \frac{s}{\sqrt{n}} \}$ является доверительным с вероятностью $P_Д$.

Замечание. Сравнение полученного интервала с интервалом,

$$(\overline{\xi} - f_p \frac{\sigma}{\sqrt{n}}, \ \overline{\xi} + f_p \frac{\sigma}{\sqrt{n}})$$

построенным при известной дисперсии (первый пример построения интервала). Видно, что в полученном интервале вместо известного значения σ фигурирует оценка s для σ, и вместо квантили f_P нормального распределения $N(0,1)$ появилась квантиль t_P распределения $S(n\text{-}1)$ Стьюдента. Отметим, что при равных доверительных вероятностях $t_P > f_P$. В табл. 2 для примера приведены некоторые значения.

Табл. 2. Сравнительные значения f_P и t_P

$P_Д$	f_P	t_P			
		n = 5	n = 10	n = 20	n = 50
0.95	1.96	2.57	2.23	2.09	2.00
0.99	2.58	4.03	3.17	2.85	2.66

6.5. Построение центральной статистики.

В общем случае центральную статистику $\varphi(\xi,a)$, закон которой известен и не зависит от параметра, можно построить, основываясь на следующей лемме.

Лемма. Пусть ξ — непрерывная случайная величина с функцией распределения $F(x)$. Тогда случайная величина

$$\eta = F(\xi) \sim R[0,1].$$

Действительно, определив функцию распределения $F_\eta(y)$ случайной величины η:

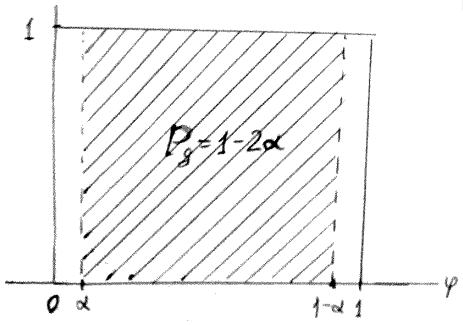
$$F_\eta(y) = P\{\eta \equiv F(\xi) < y\} = \begin{cases} 0, \text{ если } y \le 0, \\ P\{\xi < F^{-1}(y)\} = F(F^{-1}(y)) = y, \text{если } 0 < y \le 1, \\ 1, \text{ если } y > 1, \end{cases}$$

убеждаемся, что это так.

Пусть при построении доверительного интервала мы используем некоторую статистику $\zeta(\xi_1, \xi_2 \ldots \xi_n)$. Определим функцию распределения $F(z,a)$ статистики ζ (F зависит от a).

$$\zeta \sim F(z,a)$$

Случайная величина $\varphi = F(\zeta,a)$, если a есть *истинное значение* параметра, в силу леммы, распределена *равномерно на отрезке* [0, 1] при любом истинном значении a, и потому мы можем ее принять в качестве центральной статистики $\varphi(\xi,a) = F(\zeta,a)$. В качестве (9) имеем соотношение (рис.9):



$$P\{\alpha < \varphi = F(\zeta,a) < 1-\alpha\} = 1 - 2\alpha \equiv P_Д,$$

справедливое при любом значении параметра a. Разрешив два неравенства под знаком вероятности относительно a, получим доверительный интервал.

Такой подход с некоторыми изменениями можно применить и для дискретных распределений.

Рис. 9. Выбор интервала для $F(\zeta,a)$

Замечание 1. Можно рассуждать иначе. При любом фиксированном значении истинного параметра *a* определим интервал ($z_1(a)$, $z_2(a)$) так, чтобы

$$P\{z_1(a) < \zeta < z_2(a)\} = P_Д. \qquad (15)$$

Ясно, что в качестве z_1 и z_2 можно взять квантили, т.е. определить z_1 и z_2 из условий

$$F(z_1,a) = (1 - P_Д) / 2, \qquad F(z_2,a) = (1 + P_Д) / 2.$$

Если $z_1(a)$ и $z_2(a)$ монотонно ***возрастают*** по *a,* то, разрешив два неравенства под знаком P в (15) и учитывая, что $z_1(a) < z_2(a)$, получим соотношение

$$P\{ z_2^{-1}(\zeta) < a < z_1^{-1}(\zeta) \} = P_Д,$$

верное при любом *a. Я*сно, что интервал ($z_2^{-1}(\zeta)$, $z_1^{-1}(\zeta)$), определяемый двумя функциями от ζ, является доверительным с уровнем доверия $P_Д$.

Если $z_1(a)$ и $z_2(a)$ монотонно *убывают* по *a*, то доверительный интервал получим таким же образом.

При построении *односторонних границ* нужно вместо двух неравенств использовать лишь одно, согласовав знак неравенства с типом границы (верхняя или нижняя) и с характером монотонности (убывание или возрастание).

Пример 1. Пусть ξ_1, $\xi_2 \ldots \xi_n$ — независимые наблюдения над нормальной $N(m, \sigma^2)$ случайной величиной. Пусть дисперсия σ^2 известна. Построим доверительный интервал для среднего *m* с уровнем доверия $P_Д = 1 - 2\alpha$. Оценкой для *m* является статистика $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n}\xi_i$, имеющая функцию распределения $F_{\bar{\xi}}(z) = \Phi\left(\frac{z - m}{\sigma/\sqrt{n}}\right)$, где $\Phi(x)$ - функция стандартного нормального распределения. Согласно лемме, случайная величина $\Phi\left(\frac{\bar{\xi}-m}{\sigma/\sqrt{n}}\right)$ распределена равномерно на отрезке [0,1], следовательно,

$$P\left\{\alpha < \Phi\left(\frac{\bar{\xi}-m}{\sigma/\sqrt{n}}\right) < 1-\alpha\right\} = 1 - 2\alpha = P_Д$$

при любом истинном значении *m*. После разрешения двух неравенств под знаком вероятности получим

$$\mathbf{P}\left\{\bar{\xi} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha) < m < \bar{\xi} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}(1-\alpha)\right\} = P_Д,$$

где учтено, что $\Phi^{-1}(\alpha) = - \Phi^{-1}(1-\alpha)$. Обозначив $\Phi^{-1}(1-\alpha) = f_P$, получаем интервал ($\bar{\xi} - f_p\frac{\sigma}{\sqrt{n}}$, $\bar{\xi} + f_p\frac{\sigma}{\sqrt{n}}$), совпадающий с (8).

Пример 2. Пусть $\xi_1, \xi_2 \ldots \xi_n$ — независимые наблюдения над случайной величиной, распределенной равномерно на отрезке [0, *a*], где *a* — неизвестный параметр, для которого нужно построить верхнюю доверительную границу с доверительной вероятностью $P_Д = 1 - \alpha$. Оценивающей статистикой является $\zeta = \max\limits_i \xi_i$ — достаточная для *a* статистика. Определим функцию распределения случайной величины ζ:

$$F_\zeta(z, a) = P\left\{\max_i \xi_i < z\right\} = \begin{cases} 0, \text{ если } z \le 0, \\ (z/a,)^n \text{ если } 0 < z < a, \\ 1, \text{ если } z \ge a. \end{cases}$$

Случайная величина $(\zeta/a,)^n$ распределена равномерно на [0,1], и потому при любом *a*

$$P\left\{\left(\max \xi_i/a\right)^n > 1 - P_Д\right\} = P_Д = P\left\{\max_i \xi_i \Big/ \sqrt[n]{1 - P_Д} > a\right\},$$

что означает, что статистика $\max\limits_i \xi_i \Big/ \sqrt[n]{1 - P_Д}$ является верхней доверительной границей для *a* с коэффициентом доверия $P_Д$.

Верно также при любом *a*

$$P\left\{\left(\max \xi_i/a\right)^n < P_Д\right\} = P_Д = P\left\{\max_i \xi_i \Big/ \sqrt[n]{P_Д} < a\right\},$$

т.е. статистика $\max\limits_i \xi_i \Big/ \sqrt[n]{P_Д}$ является нижней доверительной границей.

Замечание 2. *Общая логика* построения доверительного множества по статистике ζ заключается в следующем. Для каждого значения параметра *a* построим множество Z(*a*) значений z случайной величины ζ вероятности $P_Д$; конечно, оно зависит от *a*, Z= Z(*a*):

$$\mathrm{P}\{\zeta \in \mathrm{Z}(a)\} = P_Д \ \forall\ a.$$

Далее для любого *z* построим множество A(z) значений параметра *a*, включив в него те значения *a*, для которых Z(*a*) содержит *z*:

$$\mathrm{A(z)} = \{a: \mathrm{z} \in \mathrm{Z}(a)\}, \text{ т.е. } a \in \mathrm{A(z)} \Leftrightarrow \mathrm{z} \in \mathrm{Z}(a),$$

и потому случайное множество A(ζ) содержит истинное значение *a* с вероятностью

$$\mathrm{P}\{\mathrm{A}(\zeta) \ni a\} = \mathrm{P}\{\zeta \in \mathrm{Z}(a)\} = P_Д \ \forall\ a.$$

§ 7. Интервалы при больших выборках

7.1. Использование асимптотической нормальности оценок.

Пусть по выборке $\xi = (\xi_1, \xi_2 \ldots \xi_n)$ оценивается неизвестный параметр *a*, и пусть оценка $\hat{a} = \hat{a}(\xi)$ асимптотически нормальна со

средним *a* и дисперсией $\sigma_n^2(a)$, зависящей от неизвестного параметр *a*. Рассмотрим нормированную погрешность

$$\varphi(\xi, a) = \frac{\hat{a}(\xi) - a}{\sigma_n(a)}. \qquad (16)$$

Эта случайная величина распределена приближенно по нормальному закону *N*(0,1) при любом значении параметра *a*. По заданному значению доверительной вероятности $P_Д$ определяем симметричный интервал ($-f_P$, f_P), который несет в себе вероятность $P_Д$ нормального *N*(0,1) распределения, и потому при любом значении параметра *a* верно приближенное соотношение:

$$P\{|\varphi(\xi, a)| < f_P\} \approx P_Д, \ \forall a .$$

Полагая монотонность φ по *a* и разрешая под знаком вероятности неравенства

$$-f_P < \frac{\hat{a}(\xi) - a}{\sigma_n(a)} < f_P, \qquad (16a)$$

относительно *a,* получим соотношение

$$P\{g_1(\hat{a}, f_P) < a < g_2(\hat{a}, f_P)\} \approx P_Д,$$

верное при любом значении параметра *a.* Это означает, что $(g_1(\hat{a}, f_P), g_2(\hat{a}, f_P))$ является доверительным интервалом коэффициентом доверия, приближенно равным $P_Д$.

Замечание. Сказанное можно обобщить. Вместо оценки $\hat{a}(\xi)$ рассматрим $\zeta = \zeta(\xi_1, \xi_2 \ldots \xi_n)$ – некоторую статистику, распределенную приближенно нормально с мат. ожиданием $m(a) = M\zeta$ и дисперсией $\sigma^2(a) = D\zeta$. Пронормировав ζ, вместо (16), получаем с.в. :

$$\varphi(\zeta, a) = \frac{\zeta - m(a)}{\sigma_n(a)} \sim N(0,1)$$

Все остальное будет справедливым, в результате получим доверительный интервал.

7.2. Примеры

A. *Доверительный интервал для вероятности*. Пусть $P(A) = a$ — неизвестная вероятность некоторого события *A*, и ξ — число появлений A в серии *n* независимых испытаний. Несмещенной оценкой для *a* является $\hat{a} = \xi/n$, которая по теореме Муавра-Лапласа при больших значениях числа испытаний *n* является асимптотически нормальной с параметрами

$$M\hat{a} = M\frac{\xi}{n} = a, \ D\hat{a} = D\frac{\xi}{n} = \frac{a(1-a)}{n} = \sigma_n^2(a),$$

и потому нормированная погрешность

$$\varphi(\hat{a}, a) = \frac{\hat{a} - a}{\sigma_n(a)} = \frac{\hat{a} - a}{\sqrt{a(1-a)}}\sqrt{n}$$

распределена приближенно по нормальному закону $N(0,1)$ при любом значении параметра a. Разрешая неравенство (16)

$$\varphi^2(\hat{a},a)=\left(\frac{\hat{a}-a}{\sqrt{a(1-a)}}\sqrt{n}\right)^2 < f_P{}^2$$

относительно a, получаем доверительный интервал $(a_1(\hat{a},f_P,n),a_2(\hat{a},f_P,n))$, где a_1, a_2 — корни соответствующего квадратного уравнения

$$a_{1,2}=\left[(\hat{a}+f_P^2/2n)\pm f_P\sqrt{\frac{\hat{a}(1-\hat{a})}{n}+\frac{f_P^2}{4n^2}}\right]\Bigg/\left(1+\frac{f_P^2}{n^2}\right).$$

При больших n формула упрощается:

$$a_{1,2}\approx \hat{a}\pm f_P\sqrt{\frac{\hat{a}(1-\hat{a})}{n}}=\hat{a}\pm f_P\sigma_n(\hat{a}).$$

При малых значениях n, когда нельзя пользоваться приближенной нормальностью, пользуются статистическими таблицами, в которых для заданных n и $P_Д$ и полученному значению оценки $\hat{a}$ указаны левый и правый концы интервала. *Вставить номограмму*

Б. Доверительный интервал для коэффициента корреляции. Пусть имеется пара случайных величин (ξ, η), для которых по имеющейся выборке (ξ_1, η_1), $(\xi_2, \eta_2)\ldots(\xi_n, \eta_n)$ нужно определить доверительный интервал для коэффициента корреляции

$$r=\frac{\operatorname{cov}(\xi,\eta)}{\sigma_\xi\sigma_\eta}=\frac{\mathrm{M}(\xi\eta)-\mathrm{M}\xi\cdot\mathrm{M}\eta}{\sqrt{\mathrm{D}\xi\cdot\mathrm{D}\eta}}.$$

напомним практический смысл:

прогноз: $\hat{\xi}(\eta)=M\left(\xi|\eta=y\right)=M\left(\xi\right)+r\frac{\sigma_\xi}{\sigma_\eta}(y-M\eta)$

Методом моментов получаем оценку для r: $\hat{r}=\dfrac{\overline{\xi\eta}-\overline{\xi}\cdot\overline{\eta}}{s_\xi\cdot s_\eta}$,

где обозначено

$$\overline{\xi}=\frac{1}{n}\sum_{i=1}^{n}\xi_i,\ \overline{\eta}=\frac{1}{n}\sum_{i=1}^{n}\eta_i,\ \overline{\xi\eta}=\frac{1}{n}\sum_{i=1}^{n}\xi_i\eta_i,\ s_\xi=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\overline{\xi}_i)^2},\ s_\eta=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\overline{\eta}_i-\overline{\eta})^2}\ .$$

Если распределение случайных величин (ξ, η) является нормальным, оценка $\hat{r}$ распределена при больших n приближенно нормально [4], причем

$$\mathrm{M}\hat{r}=r,\quad \mathrm{D}\hat{r}=\frac{(1-r^2)^2}{n-1}.$$

Этого достаточно, чтобы определить приближенный доверительный интервал. Однако, более удобным является другой способ, основанный на преобразовании z Фишера:

$$z = z(\hat{r}) = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}}.$$

Эта статистика распределена приближенно (при $n \geq 20$) по нормальному закону [4] со средним

$$m_z(r) \approx \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{r}{2(n-3)}$$

и дисперсией $\sigma^2 \approx 1 / (n - 3)$, а статистика

$$\sqrt{n-3}(z - m_z(r)) \sim N(0,1)$$

приближенно нормальна. Доверительный интервал для r получаем из неравенства

$$\left|\sqrt{n-3}(z - m_z(r))\right| < f_P,$$

где $f_Р = Q((1 + P_Д) / 2)$ — квантиль уровня $(1 + P_Д) / 2$ нормального распределения. В статистических таблицах (например, [4]) даны интервалы для заданных n, $P_Д$ и $\hat{r}$. Для примера укажем, что при $P_Д = 0{,}95$ и $\hat{r} = 0{,}8$ интервалы составляют:

(0,32; 0,95) при $n = 10$ и (0,53; 0,92) при $n = 20$.