Лекция 4. Доверительные интервалы

Буре В.М., Грауэр Л.В.

ШАД

Санкт-Петербург, 2013

Содержание

1. Доверительные интервалы
   - Общая схема построения доверительных интервалов
   - Асимптотические доверительные интервалы
   - Распределения статистик для выборок из нормальной генеральной совокупности
   - Распределение Стьюдента
   - Статистика Пирсона
   - Точные доверительные интервалы для нормальной генеральной совокупности
2. Гамма-распределение

Общая схема построения доверительных интервалов

Пусть задана генеральная совокупность ξ с функцией распределения $F_\xi(x)$. Имеется выборка $X_{[n]} = (X_1, \ldots, X_n)$ из этой генеральной совокупности и неизвестный параметр распределения $\theta \in \Theta \subset \mathbb{R}$.

Определение 1

Пусть для некоторого $\alpha \in (0,1)$ *существуют статистики* $S^-(X_{[n]}, \alpha)$ *и* $S^+(X_{[n]}, \alpha)$ *такие, что*

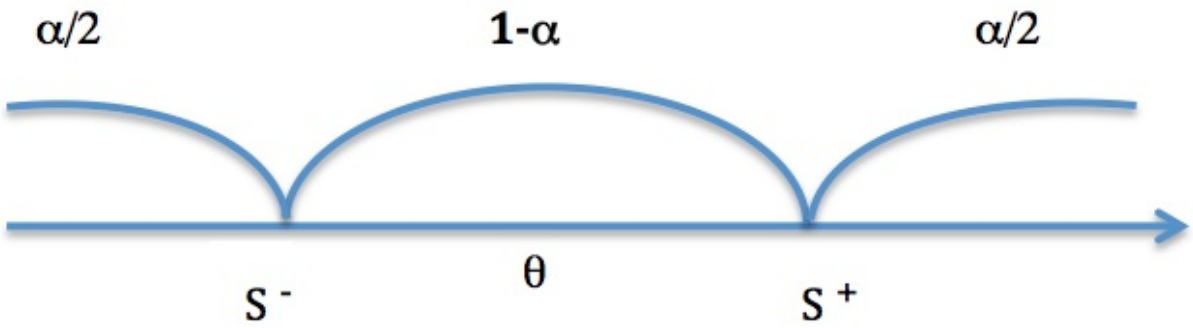
$$P\left\{S^-(X_{[n]}, \alpha) < \theta < S^+(X_{[n]}, \alpha)\right\} = 1 - \alpha,$$

тогда интервал $\left(S^-(X_{[n]}, \alpha), S^+(X_{[n]}, \alpha)\right)$ *называется доверительным интервалом для параметра* θ *с уровнем доверия* $(1 - \alpha)$.

Укажем общий метод построения доверительных интервалов, который будет использован далее.
Пусть известна статистика $Y(S(X_{[n]}), \theta)$, содержащая оцениваемый параметр θ и его точечную оценку $S(X_{[n]})$ со следующими свойствами:

1. Функция распределения $F_Y(x)$ случайной величины Y известна и не зависит от θ.
2. Функция $Y(S(X_{[n]}), \theta)$ непрерывна и строго монотонна по θ.

Зададим уровень значимости α.
Обычно доверительный интервал строят так, чтобы дополнительные интервалы $(-\inf, S^-(X_{[n]}, \alpha))$, $(S^+(X_{[n]}, \alpha), +\inf)$ накрывали θ равновероятно (с вероятностью $\alpha/2$).



Находим квантили $y_{\alpha/2}$ и $y_{1-\alpha/2}$ распределения случайной величины Y порядка $\alpha/2$ и $1-\alpha/2$ и далее получаем

$$P(y_{\alpha/2} < Y(S(X_{[n]}), \theta) < y_{1-\alpha/2}) = F(y_{1-\alpha/2}) - F(y_{\alpha/2}) = 1 - \alpha.$$

Пусть для определенности, функция $Y(S(X_{[n]}), \theta)$ строго возрастает по θ. Тогда обратная функция $Y^{-1}(y)$ для $Y(S(X_{[n]}), \theta)$ также будет строго возрастающей. Тогда неравенство

$$y_{\alpha/2} < Y(S(X_{[n]}), \theta) < y_{1-\alpha/2} \tag{1}$$

эквивалентно неравенству

$$Y^{-1}(y_{\alpha/2}) < \theta < Y^{-1}(y_{1-\alpha/2}). \tag{2}$$

Получаем доверительный интервал для θ

$$P(S^-(X_{[n]}, \alpha) < \theta < S^+(X_{[n]}, \alpha)) = 1 - \alpha,$$

где $S^-(X_{[n]}, \alpha) = Y^{-1}(y_{\alpha/2})$, $S^+(X_{[n]}, \alpha) = Y^{-1}(y_{1-\alpha/2})$.
Для случая строгого убывания $Y(S(X_{[n]}), \theta)$ по θ знаки неравенства в (1), (2) будут противоположного смысла.

Асимптотические доверительные интервалы

Определение 2

Пусть для некоторого $\alpha \in (0,1)$ *существуют статистики* $S^-(X_{[n]},\alpha)$ *и* $S^+(X_{[n]},\alpha)$ *такие, что*

$$\lim_{n\longrightarrow\infty} P\left\{S^-(X_{[n]},\alpha) < \theta < S^+(X_{[n]},\alpha)\right\} = 1-\alpha,$$

тогда интервал $\left(S^-(X_{[n]},\alpha), S^+(X_{[n]},\alpha)\right)$ *называется асимптотическим (приближенным) доверительным интервалом.*

Построение асимптотических доверительных интервалов основано на асимптотически нормальных оценках.

Предположим, что оценка $\hat{\theta} = \hat{\theta}(X_{[n]})$ является асимптотически нормальной, т. е.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow[n\to\infty]{\mathrm{d}} \varsigma \sim N(0, \sigma^2),$$

где дисперсия $\sigma^2(\theta)$ — коэффициент асимптотического рассеивания. Предположим, что функция $\sigma^2(\theta)$ непрерывна на Θ и отлична от нуля для любого $\theta \in \Theta$.

Лемма 1

Случайный вектор $(\sqrt{n}(\hat{\theta} - \theta), \hat{\theta}) \xrightarrow[n\to\infty]{d} (\zeta, \theta)$, *где* ζ *подчиняется нормальному распределению* $N(0, \sigma^2(\theta))$.

Доказательство

Покажем, что характеристическая функция случайного вектора $(\sqrt{n}(\hat{\theta} - \theta), \hat{\theta})$ удовлетворяет условию:

$$\varphi_{(\sqrt{n}(\hat{\theta}-\theta),\hat{\theta})}(t_1, t_2) \xrightarrow[n\to\infty]{} \varphi_{(\zeta,\theta)}(t_1, t_2) = Ee^{it_1\zeta + it_2\theta}.$$

Действительно,

$$\varphi_{(\sqrt{n}(\hat{\theta}-\theta),\hat{\theta})}(t_1,t_2) = Ee^{it_1\sqrt{n}(\hat{\theta}-\theta)+i(t_2\hat{\theta}\pm\theta t_2)} =$$

$$= Ee^{i\sqrt{n}(\hat{\theta}-\theta)(t_1+\frac{t_2}{\sqrt{n}})}e^{it_2\theta} = e^{it_2\theta}\varphi_{\sqrt{n}(\hat{\theta}-\theta)}(t_1+\frac{t_2}{\sqrt{n}}) =$$

$$= e^{it_2\theta}\left(\left\{\varphi_{\sqrt{n}(\hat{\theta}-\theta)}(t_1+\frac{t_2}{\sqrt{n}})-\varphi_\zeta(t_1+\frac{t_2}{\sqrt{n}})\right\}+\varphi_\zeta(t_1+\frac{t_2}{\sqrt{n}})\right).$$

При этом, $\varphi_\zeta(t_1+t_2/\sqrt{n}) \xrightarrow[n\to\infty]{} \varphi_\zeta(t_1)$, так как любая характеристическая функция равномерно непрерывна.
Имеет место сходимость:

$$\varphi_{\sqrt{n}(\hat{\theta}-\theta)}(t_1+\frac{t_2}{\sqrt{n}})-\varphi_\zeta(t_1+\frac{t_2}{\sqrt{n}}) \xrightarrow[n\to\infty]{} 0,$$

так как при любом t: $\varphi_{\sqrt{n}(\hat{\theta}-\theta)}(t) \to \varphi_\zeta(t)$, и сходимость равномерна на любом конечном промежутке.

Следовательно, выполняется сходимость:

$$\varphi_{(\sqrt{n}(\hat{\theta}-\theta),\hat{\theta})}(t_1,t_2) \xrightarrow[n\to\infty]{} e^{it_2\theta}\varphi_{\zeta}(t_1) = Ee^{it_2\theta+i\zeta t_1} = \varphi_{(\zeta,\theta)}(t_1,t_2).$$

Лемма доказана.

Рассмотрим функцию от двух переменных $H(x_1,x_2) = x_1/\sigma(x_2)$, она непрерывна на $\mathbb{R}\times\Theta$. Случайный вектор $(\zeta,\theta)^T \in \mathbb{R}\times\Theta$, следовательно, можем воспользоваться теоремой непрерывности:

$$H(\sqrt{n}(\hat{\theta}_n-\theta),\hat{\theta}_n) \xrightarrow[n\to\infty]{\mathrm{d}} H(\zeta,\theta) = \frac{\zeta}{\sigma(\theta)} \sim N(0,1).$$

Таким образом, имеет место сходимость:

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma(\hat{\theta})} \xrightarrow[n\to\infty]{\mathrm{d}} \eta \sim N(0,1).$$

Тогда справедливо следующее соотношение:

$$P\{-z_{1-\frac{\alpha}{2}} < \frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma(\hat{\theta})} < z_{1-\frac{\alpha}{2}}\} \xrightarrow[n\longrightarrow\infty]{} 1-\alpha = \frac{1}{\sqrt{2\pi}} \int\limits_{-z_{1-\frac{\alpha}{2}}}^{z_{1-\frac{\alpha}{2}}} e^{-\frac{y^2}{2}}\, dy,$$

где $z_{1-\frac{\alpha}{2}}$ — квантиль стандартного нормального распределения уровня $1-\alpha/2$, то есть, $F(z_{1-\frac{\alpha}{2}}) = 1-\alpha/2$, где $F(x)$ — функция стандартного нормального распределения.
Получаем асимптотический доверительный интервал с уровнем доверия $1-\alpha$:

$$P\{\hat{\theta} - z_{1-\frac{\alpha}{2}}\frac{\sigma(\hat{\theta})}{\sqrt{n}} < \theta < \hat{\theta} + z_{1-\frac{\alpha}{2}}\frac{\sigma(\hat{\theta})}{\sqrt{n}}\} \approx 1-\alpha.$$

Ширина доверительного интервала характеризует точность интервальной оценки.

Пример 1

Рассмотрим схему Бернулли, в которой n испытаний. Пусть m — число успехов. Выборка $X_{[n]} = (a_1, \ldots, a_n)$ *состоит из последовательности нулей и единиц, тогда функция правдоподобия имеет вид:*

$$L(X_{[n]}, p) = p^m q^{n-m}, \quad p \in \Theta = (0, 1),$$

где m — число единиц в выборке. Логарифмическая функция правдоподобия имеет вид:

$$\ln L = m \ln p + (n - m) \ln(1 - p).$$

Найдем оценку максимального правдоподобия:

$$\frac{\partial \ln L}{\partial p} = \frac{m}{p} - \frac{n - m}{1 - p} = \frac{m - mp - np + mp}{p(1 - p)} = 0.$$

Следовательно, получаем оценку:

$$\hat{p} = \frac{m}{n}.$$

Убеждаемся, что $\hat{p}$ максимизирует функцию правдоподобия:

$$\frac{\partial^2 \ln L}{\partial p^2} = -\frac{m}{p^2} - \frac{n-m}{(1-p)^2} < 0.$$

Следовательно, $\hat{p} = \frac{m}{n}$ — точка максимума или оценка по методу максимального правдоподобия.

Нетрудно показать, что оценка $\hat{p}$ асимптотически нормальна:

$$\sqrt{n}\left(\frac{m}{n} - p\right) = \frac{m - np}{\sqrt{n}} = \frac{\sum\limits_{i=1}^{n} \xi_i - np}{\sqrt{n}} =$$

$$= \frac{\sum\limits_{i=1}^{n} (\xi_i - p)}{\sqrt{n}} \xrightarrow[n\to\infty]{\mathrm{d}} \zeta \sim N(0, pq),$$

где $P\{\xi_i = 1\} = p$, $P\{\xi_i = 0\} = q = 1 - p$, $\sigma^2 = pq = p(1-p)$.

Воспользуемся доказанным утверждением:

$$\frac{\sqrt{n}(\hat{\theta}-\theta)}{\sigma(\hat{\theta})} \xrightarrow[n\to\infty]{\mathrm{d}} \eta \sim N(0,1).$$

Тогда имеет место сходимость:

$$\frac{\sqrt{n}(\frac{m}{n}-p)}{\sqrt{\frac{m}{n}(1-\frac{m}{n})}} \xrightarrow[n\to\infty]{\mathrm{d}} \eta \sim N(0,1).$$

Следуя приведенным выше рассуждениям, получаем доверительный интервал с уровнем доверия $1-\alpha$ для вероятности p:

$$\left(\frac{m}{n} - z_{1-\frac{\alpha}{2}}\frac{\sqrt{\frac{m}{n}(1-\frac{m}{n})}}{\sqrt{n}},\ \frac{m}{n} + z_{1-\frac{\alpha}{2}}\frac{\sqrt{\frac{m}{n}(1-\frac{m}{n})}}{\sqrt{n}}\right)$$

Распределения статистик для выборок из нормальной генеральной совокупности

Пусть имеется генеральная совокупность $\xi \sim N(a, \sigma^2)$ и выборка $X_{[n]}$ из этой генеральной совокупности. Если ξ — гауссова случайная величина, то функция плотности ее распределения имеет вид:

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}}, \quad x \in \mathbb{R},$$

а если ξ — гауссов случайный вектор, то

$$f_\xi(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-a)^T\Sigma^{-1}(x-a)}, \quad x \in \mathbb{R}^m.$$

В первом случае предполагается, что $\sigma^2 > 0$, а во-втором, что $\det|\Sigma| \neq 0$.

Любому распределению взаимно однозначно соответствует характеристическая функция. С помощью метода характеристических функций легко получить, что компоненты гауссова случайного вектора независимы тогда и только тогда, когда ковариационная матрица Σ диагональна или, другими словами, когда равны нулю попарные ковариации всех компонент.

Лемма 2

Пусть ζ — случайный вектор размерности m, подчиняющийся многомерному нормальному распределению $N(a, \Sigma)$, пусть A — любая матрица размерности $n \times m$, b — вектор размерности $n \times 1$. Тогда вектор $\eta = A\zeta + b$ подчиняется нормальному распределению $N\left(Aa + b; A\Sigma A^T\right)$.

Доказательство

Рассмотрим

$$\varphi_\eta(t) = Ee^{it^T\eta} = e^{it^Tb}Ee^{i(t^TA)\zeta} = e^{it^Tb}\varphi_\zeta(t^TA) = \\ = Ee^{it^T(Aa+b)-\frac{1}{2}t^TA\Sigma A^Tt},$$

где $Aa + b$ — математическое ожидание, $A\Sigma A^T$ — ковариационная матрица случайного вектора η.

Лемма доказана.

Лемма 3

Пусть $\xi \sim N\left(0, \sigma^2 E_n\right)$ *и* $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$ *— среднее арифметическое. Тогда* $\bar{\xi}$ *и вектор* $\left(\xi_1 - \bar{\xi}, \ldots, \xi_n - \bar{\xi}\right)$ *взаимно независимы.*

Доказательство

Возьмем любой элемент вектора, например, $\xi_k - \bar{\xi}$, и проверим его независимость с $\bar{\xi}$. Рассмотрим разность:

$$\xi_k - \bar{\xi} = \xi_k - \frac{1}{n}\sum_{i=1}^{n} \xi_i = -\frac{1}{n}\xi_1 - \ldots - \frac{1}{n}\xi_{k-1} + \frac{n-1}{n}\xi_k -$$

$$- \frac{1}{n}\xi_{k+1} - \ldots - \frac{1}{n}\xi_n = \left(-\frac{1}{n}, \ldots, -\frac{1}{n}, \frac{n-1}{n}, -\frac{1}{n}, \ldots - \frac{1}{n}\right)\xi,$$

$$\bar{\xi} = \frac{1}{n}\xi_1 + \ldots + \frac{1}{n}\xi_n = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)\xi,$$

$$\begin{pmatrix} -\frac{1}{n}, \ldots, -\frac{1}{n}, \frac{n-1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n} \\ \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \end{pmatrix}\xi = \begin{pmatrix} \xi_k - \bar{\xi} \\ \bar{\xi} \end{pmatrix}.$$

Тогда по лемме 2 получаем:

$$\begin{pmatrix} \xi_k - \bar{\xi} \\ \bar{\xi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right),$$

где $\sigma_1^2 = \frac{n-1}{n}\sigma^2$, $\sigma_2^2 = \frac{1}{n}\sigma^2$, внедиагональные элементы равны нулю, поскольку

$$\sigma^2 \left(-\frac{1}{n}, \ldots, -\frac{1}{n}, \frac{n-1}{n}, -\frac{1}{n}, \ldots, -\frac{1}{n} \right) \left(\frac{1}{n}, \ldots, \frac{1}{n} \right)^T =$$

$$= \sigma^2 \frac{-(n-1) + (n-1)}{n^2} = 0.$$

Лемма доказана

Лемма 4

Пусть $\xi = (\xi_1, \ldots, \xi_m)^T \sim N(0, E_m)$, $CC^T = C^T C = E_m$ *и* $\eta = C\xi$.
Тогда $\tau = \sum_{k=1}^{m} \xi_k^2 - \eta_1^2 - \ldots - \eta_r^2$ *подчиняется распределению хи-квадрат с* $m - r$ *степенями свободы, и случайные величины* $\eta_1, \ldots, \eta_r$ *взаимно независимы с* τ.

Доказательство
Из леммы 2 следует, что $\eta \sim N(0, E_{m\times m})$. Как легко видеть, имеет место равенство:

$$\sum_{i=1}^{m} \xi_i^2 = \xi^T \xi = \xi^T C^T C \xi = \eta^T \eta = \sum_{i=1}^{m} \eta_i^2.$$

Следовательно, справедливо равенство:

$$\sum_{i=1}^{m} \xi_i^2 - \eta_1^2 - \ldots - \eta_r^2 = \sum_{j=r+1}^{m} \eta_j^2,$$

полученное равенство доказывает лемму.

Распределение Стьюдента

Определение 3

Пусть заданы случайные величины $\zeta \sim N(0,1)$ *и* $\tau_k \sim \chi^2_k$. *Пусть случайные величины* ζ *и* τ_k *взаимно независимы. Распределение случайной величины*

$$\xi = \frac{\zeta}{\sqrt{\frac{\tau_k}{k}}}$$

называется распределением Стьюдента с k *степенями свободы и обозначается через* T_k.

Замечание 1

Очевидно, что если $\zeta \sim N(0,1)$ *и* $\zeta_1,\ldots,\zeta_k$ *— взаимно независимые случайные величины, подчиняющиеся стандартному нормальному распределению, независимые с* ζ*, тогда*

$$\frac{\zeta}{\sqrt{\frac{\zeta_1^2+\ldots+\zeta_k^2}{k}}} \sim T_k.$$

Лемма 5

Пусть $\zeta = \eta/\xi$*, где* η*,* ξ *— взаимно независимые случайные величины. Пусть* $\xi \overset{п.н.}{>} 0$*,* $f_\xi(x)$*,* $f_\eta(y)$ *— плотности распределения* ξ *и* η *соответственно, тогда плотность распределения* $f_\zeta(z)$ *дроби* ζ *имеет вид:*

$$f_\zeta(z) = \int_0^\infty x f_\xi(x) f_\eta(zx) dx.$$

Доказательство

Найдем функцию распределения случайной величины ζ:

$$F_\zeta(z) = P\{\zeta \le z\} = \int\limits_{\{(x,y):\frac{y}{x}\le z, x>0\}} f_\xi(x) f_\eta(y)\,dxdy =$$

$$= \int\limits_0^\infty \int\limits_{-\infty}^{zx} f_\xi(x) f_\eta(y)\,dydx = \int\limits_0^\infty f_\xi(x) \left(\int\limits_{-\infty}^{z} f_\eta(xs)x\,ds \right) dx.$$

Положим $s = y/x$, $dy = xds$. Меняем порядок интегрирования:

$$F_\zeta(z) = \int\limits_{-\infty}^{z} \left(\int\limits_0^\infty x f_\xi(x) f_\eta(xs)\,dx \right) ds.$$

Полученное равенство доказывает лемму.

Лемма 6

Пусть $\eta \sim N(0,1)$, $\xi = \sqrt{\tau}$, $\tau \sim \chi^2_k$. *Пусть случайные величины* η *и* τ *взаимно независимы. Тогда случайная величина* $\zeta = \eta/\xi$ *имеет плотность распределения следующего вида:*

$$f_\zeta(z) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi}\Gamma(\frac{k}{2})} \frac{1}{(1+z^2)^{\frac{k+1}{2}}}$$

Доказательство

Распределение хи-квадрат с k степенями свободы представляет собой гамма-распределение с параметрами формы $k/2$ и масштаба $1/2$:

$$f_\tau(x) = \begin{cases} (\frac{1}{2})^{\frac{k}{2}} \frac{x^{\frac{k}{2}-1}}{\Gamma(\frac{k}{2})} e^{-\frac{x}{2}}, & x > 0; \\ 0, & x \le 0. \end{cases}$$

Как легко заметить, имеет место равенство:

$$P\{\sqrt{\tau} \le x\} = P\{\tau \le x^2\},$$

тогда

$$f_{\sqrt{\tau}}(x) = 2xf_\tau(x^2) = \begin{cases} \frac{1}{2^{\frac{k}{2}-1}}\frac{x^{k-1}}{\Gamma(\frac{k}{2})}e^{-\frac{x^2}{2}}, & x > 0; \\ 0, & x \le 0. \end{cases}$$

Следовательно, имеют место равенства:

$$f_\zeta(z) = \int\limits_0^\infty xf_\xi(x)f_\eta(zx)dx = \frac{1}{2^{\frac{k}{2}-1}\Gamma(\frac{k}{2})\sqrt{2\pi}}\int\limits_0^\infty x^k e^{-\frac{x^2}{2}}e^{-\frac{z^2x^2}{2}}dx =$$

$$= \frac{1}{2^{\frac{k-1}{2}}\Gamma(\frac{k}{2})\sqrt{\pi}}\int\limits_0^\infty x^k e^{-\frac{x^2}{2}(z^2+1)}dx =$$

$$= \frac{1}{\sqrt{\pi}\Gamma(\frac{k}{2})}\frac{1}{(z^2+1)^{\frac{k+1}{2}}}\int\limits_0^\infty u^{\frac{k+1}{2}-1}e^{-u}du,$$

где $u = x^2(z^2+1)/2$, $xdx = du/(z^2+1)$, причем,
$\int\limits_0^\infty u^{\frac{k+1}{2}-1}e^{-u}du = \Gamma\left(\frac{k+1}{2}\right)$.

Следствие 1

Плотность распределения Стьюдента с k степенями свободы имеет вид:

$$f(z) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\Gamma(\frac{k}{2})} \frac{1}{(1+z^2/k)^{\frac{k+1}{2}}}. \quad (3)$$

Доказательство

Для доказательства достаточно заметить, что случайная величина $\sqrt{k}\zeta$ подчиняется распределению Стьюдента с k степенями свободы и $f_{\sqrt{k}\zeta}(z) = \frac{1}{\sqrt{k}} f_\zeta(\frac{z}{\sqrt{k}})$.

Замечание 2

Можно показать, что для плотности $f(z)$ из выражения (3) имеет место сходимость:

$$f(z) \xrightarrow[k\to\infty]{} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

Статистика Пирсона

Пусть задана генеральная совокупность ξ с функцией распределения F_ξ и выборка $X_{[n]} = (X_1, \ldots, X_n)$. Разобьем числовую ось на r непересекающихся интервалов: $-\infty = a_0 < a_1 < \ldots < a_r = \infty$. Обозначим через $\Delta_1 = (-\infty, a_1]$, $\Delta_2 = (a_1, a_2]$, ..., $\Delta_r = (a_{r-1}, \infty)$. Пусть $p_i = F_\xi(a_i) - F_\xi(a_{i-1})$ — вероятность того, что случайная величина ξ попадет в интервал Δ_i, $\sum_{i=1}^r p_i = 1$. Пусть n_i — количество элементов выборки $X_{[n]}$, попавших в Δ_i. Определим статистику χ^2 следующим образом:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}, \tag{4}$$

где n_i — частота (количество элементов выборки, попавших в Δ_i), np_i — ожидаемое количество наблюдений в интервале Δ_i.

Определение 4

Статистики вида 4 называются статистиками χ^2 или статистиками Пирсона.

Покажем, что статистика Пирсона может быть преобразована к статистике первого типа:

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} = n \sum_{i=1}^{r} \left(\frac{1}{\sqrt{p_i}} \frac{n_i}{n} - \sqrt{p_i} \right)^2 =$$

$$= n \sum_{i=1}^{r} \left(\sum_{j=1}^{n} \frac{1}{\sqrt{p_i}} I\{X_j \in \Delta_i\} - \sqrt{p_i} \right)^2 . \quad (5)$$

Рассмотрим статистику первого типа

$$S(X_{[n]}) = h(\frac{1}{n}\sum_{j=1}^{n} g(X_j)),$$

где в качестве h возьмем функцию $h(t_1, \ldots, t_r) = = \sum_{i=1}^{r} \left(t_i - \sqrt{p_i}\right)^2$, $g(x) = (g_1(x), \ldots, g_r(x))$ и $g_i(x) = \frac{1}{\sqrt{p_i}} I\{x \in \Delta_i\}$, $i = 1, \ldots, r$.
Получаем, что

$$\frac{1}{n}\sum_{j=1}^{n} g(X_j) = \left(\frac{1}{\sqrt{p_1}}\frac{n_1}{n}, \ldots, \frac{1}{\sqrt{p_r}}\frac{n_r}{n}\right).$$

Таким образом, статистика χ^2 представляет собой произведение константы n на статистику первого типа. Следовательно, можно воспользоваться теоремой 8 (Л2) о предельном распределении статистик первого типа.

Очевидно, что $a = (a_1, \ldots, a_r) = (Eg_1(\xi), \ldots, Eg_r(\xi)) = (\sqrt{p_1}, \ldots, \sqrt{p_r})$, при этом $h'(a) = 0$, $\frac{1}{2}h''(a) = E_r$, где E_r — единичная матрица порядка r. Из теоремы 8 (Л2) получаем, что

$$\chi^2(X_{[n]}) \xrightarrow[n\to\infty]{\mathrm{d}} \zeta^T\zeta = \zeta_1^2 + \ldots + \zeta_r^2,$$

где $\zeta \sim N(0, Dg(\xi))$. Вычислим ковариационную матрицу $Dg(\xi)$:

$$Dg(\xi) = E\left\{g(\xi)g^T(\xi)\right\} - Eg(\xi)\left(Eg(\xi)\right)^T,$$

после дополнительных преобразований получаем

$$Dg(\xi) = E_r - (\sqrt{p_1}, \ldots, \sqrt{p_r})^T(\sqrt{p_1}, \ldots, \sqrt{p_r}).$$

Пусть C — ортонормированная матрица $CC^T = C^T C = E_r$.
Зафиксируем первую строку $c_1 = (\sqrt{p_1}, \ldots, \sqrt{p_r})$ матрицы C.
Остальные строки будем искать методом ортогонализации Грамма-Шмидта.
Случайный вектор $\eta = C\zeta$ подчиняется многомерному нормальному распределению $N(0, CDg(\xi)C^T)$, при этом из выбора матрицы C следует, что $D\eta_1 = 0$, кроме того, $E\eta_1 = 0$, следовательно, $\eta_1 \overset{\text{п.н.}}{=} 0$.

$$CDg(\xi)C^T = \begin{pmatrix} 0 & 00\ldots0 \\ 0 & \\ \ldots & E_{r-1} \\ 0 & \end{pmatrix}.$$

Как легко видеть,
$\eta^T\eta = \zeta^T C^T C\zeta = \zeta^T\zeta = \eta_1^2 + \eta_2^2 + \ldots + \eta_r^2 = \zeta_1^2 + \zeta_2^2 + \ldots + \zeta_r^2$.

Следовательно, распределения сумм одинаковы, поэтому

$$\chi^2 \xrightarrow[n\to\infty]{\mathrm{d}} \eta_1^2 + \ldots + \eta_r^2,$$

но $\eta_1^2 + \ldots + \eta_r^2 \overset{\text{п.н.}}{=} \eta_2^2 + \ldots + \eta_r^2$, тогда

$$\chi^2 \xrightarrow[n\to\infty]{\mathrm{d}} \eta_2^2 + \ldots + \eta_r^2,$$

где $\eta_2, \ldots, \eta_r$ — взаимно независимые одинаково распределенные случайные величины, $\eta_i \sim N(0,1)$, $i = 2, \ldots, r$, $\eta_1 \overset{\text{п.н.}}{=} 0$.

Определение 5

Пусть $\delta_1, \ldots, \delta_k$ — взаимно независимые одинаково распределенные стандартные гауссовы случайные величины, тогда распределение случайной величины $\delta_1^2 + \ldots + \delta_k^2$, называется распределением χ^2 с k степенями свободы (или распределением Пирсона с k степенями свободы).

Проведенные рассуждения доказывают следующую теорему.

Теорема 1

Статистика χ^2, определяемая равенством (4), асимптотически распределена по закону хи-квадрат с $r-1$ степенью свободы, то есть

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \xrightarrow[n\to\infty]{d} \tau,$$

где τ подчиняется распределению хи-квадрат с $r-1$ степенью свободы.

Замечание 3

Нетрудно заметить, что к статистике χ^2, определяемой формулой (4), можно прийти, исходя из генеральной совокупности, подчиняющейся полиномиальному распределению с r возможными исходами, где вероятности p_1, p_2,...,p_r представляют собой вероятности появления соответствующих исходов ($\sum_{i=1}^{r} p_i = 1$), n — число испытаний, n_1 — количество появлений первого исхода, n_2 — количество появлений второго исхода, ..., n_r — количество появлений исхода с номером r, $\sum_{i=1}^{r} n_i = n$. В статистическом эксперименте непосредственно наблюдается выборка частот $(n_1, n_2, \ldots, n_r)$.

Утверждение теоремы 1 сохраняется и для рассматриваемого полиномиального распределения.

Точные доверительные интервалы для нормальной генеральной совокупности

Теорема 2

Пусть задана выборка $X_{[n]}$ *из генеральной совокупности* $\xi \sim N(a, \sigma^2)$. *Справедливы следующие утверждения:*

1. *Статистика* $\frac{\bar{X}-a}{\sigma}\sqrt{n}$ *подчиняется стандартному нормальному распределению.*
2. *Если* $\tilde{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, *тогда статистика* $\frac{\bar{X}-a}{\tilde{s}}\sqrt{n}$ *подчиняется распределению Стьюдента с* $n-1$ *степенью свободы.*
3. *Статистика* $\frac{(n-1)\tilde{s}^2}{\sigma^2}$ *подчиняется распределению хи-квадрат с* $n-1$ *степенью свободы.*
4. *Если* $s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$, *тогда статистика* $\frac{ns^2}{\sigma^2}$ *подчиняется распределению хи-квадрат с* n *степенями свободы.*

Тогда справедливо равенство:

$$\frac{(n-1)\tilde{s}^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - a}{\sigma}\right)^2 - \left(\sum_{j=1}^{n}\frac{1}{\sqrt{n}}\frac{X_j - a}{\sigma}\right)^2.$$

Введем обозначение:

$$\delta = \begin{pmatrix} \frac{X_1-a}{\sigma} \\ \dots \\ \frac{X_n-a}{\sigma} \end{pmatrix},$$

при этом δ подчиняется многомерному нормальному распределению с параметрами $(0, E_n)$.
Рассмотрим строку $(1/\sqrt{n}, \dots, 1/\sqrt{n}) = C_1$. При этом, нетрудно заметить, что $C_1 C_1^T = 1$, тогда методом ортогонализации Грама-Шмидта последовательно получим $n-1$ строку $C_2, \dots, C_n$. Строки будут ортогональными и нормированными. Составим матрицу:

$$C = \begin{pmatrix} C_1 \\ \dots \\ C_n \end{pmatrix}.$$

Рассмотрим преобразование $C\delta = \eta$, из лемм 2 и 4 следует:

$$\frac{(n-1)\tilde{s}^2}{\sigma^2} = \sum_{i=1}^{n} \delta_1^2 - \eta_1^2 \sim \chi^2_{n-1},$$

где $\delta_i = \frac{X_i - a}{\sigma}$, $\eta_1 = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} \frac{X_j - a}{\sigma} = \sum_{j=1}^{n} \frac{1}{\sqrt{n}} \delta_i = C_1\delta$. Утверждение 3 теоремы доказано.

В лемме 3 было доказано, что $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ и $\bar{X} - a$ взаимно независимы. Рассмотрим дробь Стьюдента:

$$\frac{\frac{\bar{X} - a}{\sigma}\sqrt{n}}{\sqrt{\frac{1}{n-1}\frac{(n-1)\tilde{s}^2}{\sigma^2}}} = \frac{\bar{X} - a}{\tilde{s}}\sqrt{n} \sim T_{n-1}.$$

Таким образом, доказано утверждение 2 теоремы.

Утверждение 4 следует из определения:

$$\frac{ns^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - a)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - a}{\sigma}\right)^2 \sim \chi^2_n.$$

Теорема доказана.

Точные доверительные интервалы для нормальной генеральной совокупности можно построить по следующим правилам:

- Если a неизвестно, σ^2 известно, тогда

$$P\left\{\bar{X} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\varepsilon}{2}} < a < \bar{X} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\varepsilon}{2}}\right\} = 1 - \varepsilon,$$

  где $z_{1-\frac{\varepsilon}{2}}$ — квантиль стандартного нормального распределения.

- Если a неизвестно, σ^2 неизвестно, то доверительный интервал для a будет иметь вид:

$$P\left\{\bar{X} - \frac{\tilde{s}}{\sqrt{n}} t_{1-\frac{\varepsilon}{2}} < a < \bar{X} + \frac{\tilde{s}}{\sqrt{n}} t_{1-\frac{\varepsilon}{2}}\right\} = 1 - \varepsilon,$$

  где $t_{1-\frac{\varepsilon}{2}}$ — квантиль распределения Стьюдента с $n-1$ степенью свободы.

- Если a неизвестно, σ^2 неизвестно, необходимо построить доверительный интервал для σ^2. Пусть $\varepsilon = \varepsilon_1 + \varepsilon_2$. Пусть $u_{1-\varepsilon_2}$ — квантиль распределения хи-квадрат с $n-1$ степенью свободы уровня $1-\varepsilon_2$, u_{ε_1} — квантиль распределения хи-квадрат с $n-1$ степенью свободы уровня ε_1, тогда доверительный интервал для σ^2 будет следующим:

$$P\left\{\frac{(n-1)\tilde{s}^2}{u_{1-\varepsilon_2}} < \sigma^2 < \frac{(n-1)\tilde{s}^2}{u_{\varepsilon_1}}\right\} = 1-\varepsilon.$$

- Если a известно, σ^2 неизвестно, тогда доверительный интервал строится также, как и в случае 3, только в качестве статистики рассматривается статистика ns^2/σ^2 из пункта 4 теоремы 2. Пусть v_{ε_1} — квантиль распределения хи-квадрат с n степенями свободы уровня ε_1, $v_{1-\varepsilon_2}$ — квантиль распределения хи-квадрат с n степенями свободы уровня $1-\varepsilon_2$, тогда доверительный интервал для σ^2 будет

$$P\left\{\frac{ns^2}{v_{1-\varepsilon_2}} < \sigma^2 < \frac{ns^2}{v_{\varepsilon_1}}\right\} = 1-\varepsilon,$$

здесь $\varepsilon = \varepsilon_1 + \varepsilon_2$

Гамма-распределение

Плотность распределения случайной величины ξ, соответствующая стандартному гамма-распределению с параметром формы p, определяется формулой:

$$f_\xi(x) = \begin{cases} \frac{x^{p-1}}{\Gamma(p)} e^{-x}, & x > 0 \\ 0, & x \leqslant 0. \end{cases}$$

Нетрудно заметить, что $f_\xi(x) \geq 0$ и $\int_{-\infty}^{+\infty} f_\xi(x)dx = 1$.
Гамма-функция определяется следующим образом

$$\Gamma(p) = \int_0^\infty x^{p-1} e^{-x} dx, \qquad p > 0.$$

Свойства гамма-функции:

- Справедливы равенства:

$$\Gamma(1) = 1, \Gamma(2) = 1,$$

$$\Gamma\left(\frac{1}{2}\right) = \int_0^{+\infty} x^{-\frac{1}{2}} e^{-x} dx = 2 \int_0^{+\infty} e^{-x} d(x^{\frac{1}{2}}) = 2 \int_0^{+\infty} e^{-y^2} dy =$$

$$= \int_{-\infty}^{+\infty} e^{-y^2} dy = \left\{ \left(\int_{-\infty}^{+\infty} e^{-y^2} dy \right)^2 \right\}^{\frac{1}{2}} = \left\{ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-y^2} e^{-x^2} dxdy \right\}^{\frac{1}{2}} = \sqrt{\pi},$$

- При $p > 1$, интегрируя по частям, нетрудно получить равенство:

$$\Gamma(p) = (p-1)\Gamma(p-1).$$

Если $n \in \mathbb{N}$, то $\Gamma(n) = (n-1)!$.

Рассмотрим случайную величину $\eta = \xi/\lambda$, параметр $\lambda > 0$, нетрудно получить выражение для плотности распределения, соответствующего гамма-распределению $G(\lambda, p)$ с параметром формы p и параметром масштаба λ:

$$f_\eta(x) = \begin{cases} \frac{\lambda^p x^{p-1}}{\Gamma(p)} e^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0. \end{cases} \tag{6}$$

Если в формуле 6 положить $\lambda = 1$, то получим плотность стандартного гамма-распределения, а если в формуле (6) положить $p = 1$, то получим плотность экспоненциального распределения.

Лемма 7

Пусть δ — случайная величина, подчиняющаяся стандартному нормальному распределению, $\delta \sim N(0, 1)$, тогда случайная величина δ^2 подчиняется гамма-распределению с параметрами $p = 1/2$, $\lambda = 1/2$.

Доказательство

Пусть $\delta \sim N(0,1)$, то есть, $f_\delta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. Функцию распределения случайной величины δ обозначим через

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{y} e^{-\frac{t^2}{2}}\, dt.$$

Найдем функцию распределения случайной величины δ^2:

$$P\{\delta^2 \le y\} = P(-\sqrt{y} \le \delta \le \sqrt{y}) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y}),$$

дифференцируя, найдем плотность распределения для $y > 0$:

$$f_{\delta^2}(y) = \frac{1}{2}\frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} + \frac{1}{2}\frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}} e^{-\frac{y}{2}} =$$

$$= (\frac{1}{2})^{\frac{1}{2}} y^{-\frac{1}{2}} \frac{1}{\sqrt{\pi}} e^{-\frac{y}{2}} = \begin{cases} \frac{(\frac{1}{2})^{\frac{1}{2}} y^{-\frac{1}{2}}}{\Gamma(\frac{1}{2})} e^{-\frac{y}{2}}, & y > 0; \\ 0, & y < 0. \end{cases}$$

Следовательно, случайная величина δ^2 подчиняется гамма-распределению с параметрами $p = 1/2$, $\lambda = 1/2$.

Лемма 8

Пусть заданы взаимно независимые случайные величины $\xi_1 \sim G(\lambda, p_1)$, $\xi_2 \sim G(\lambda, p_2)$, *тогда* $\xi = \xi_1 + \xi_2 \sim G(\lambda, p_1 + p_2)$.

Доказательство

Очевидно, что

$$P\{\xi \leq y\} = \iint\limits_{\{(x_1,x_2)|x_1+x_2\leq y\}} f_{\xi_1}(x_1) f_{\xi_2}(x_2) dx_1 dx_2 =$$

$$= \int\limits_{-\infty}^{+\infty} f_{\xi_1}(x_1) \int\limits_{-\infty}^{y-x_1} f_{\xi_2}(x_2) dx_2 dx_1.$$

Сделаем замену переменных: $u = x_1 + x_2$, $du = dx_2$, тогда

$$P\{\xi \leq y\} = \int\limits_{-\infty}^{+\infty} f_{\xi_1}(x_1) \int\limits_{-\infty}^{y} f_{\xi_2}(u - x_1) du dx_1 =$$

$$= \int\limits_{-\infty}^{y} \int\limits_{-\infty}^{+\infty} f_{\xi_1}(x_1) f_{\xi_2}(u - x_1) dx_1 du.$$

Получили формулу свертки для плотности суммы двух независимых случайных величин:

$$f_\xi(u) = \int\limits_{-\infty}^{+\infty} f_{\xi_1}(x_1) f_{\xi_2}(u - x_1) dx_1 = \int\limits_{0}^{u} f_{\xi_1}(x_1) f_{\xi_2}(u - x_1) dx_1 =$$

$$= \int\limits_{0}^{u} \frac{\lambda^{p_1+p_2}}{\Gamma(p_1)\Gamma(p_2)} x_1^{p_1-1} (u - x_1)^{p_2-1} e^{-\lambda x_1} e^{-\lambda(u-x_1)} dx_1 =$$

$$= \frac{\lambda^{p_1+p_2} e^{-\lambda u}}{\Gamma(p_1)\Gamma(p_2)} \int\limits_{0}^{u} x_1^{p_1-1} (u - x_1)^{p_2-1} dx_1.$$

Сделаем замену переменных под знаком интеграла: $x_1 = su$, $s \in (0,1)$, $dx_1 = uds$, тогда

$$f_\xi(u) = \frac{\lambda^{p_1+p_2} e^{-\lambda u}}{\Gamma(p_1)\Gamma(p_2)} u^{p_1+p_2-1} \int_0^1 s^{p_1-1}(1-s)^{p_2-1} ds =$$

$$= \begin{cases} cu^{p_1+p_2-1} e^{-\lambda u} \lambda^{p_1+p_2}, & u > 0; \\ 0, & u \le 0, \end{cases}$$

где $c = (\int_0^1 s^{p_1-1}(1-s)^{p_2-1} ds)/(\Gamma(p_1)\Gamma(p_2))$. Найдем c из условия нормировки:

$$c \int_0^{+\infty} \lambda^{p_1+p_2} u^{p_1+p_2-1} e^{-\lambda u} du = 1.$$

Сделаем замены переменных: $x = \lambda u$, $u = x/\lambda$, $du = dx/\lambda$, тогда

$$c \int_0^{+\infty} x^{p_1+p_2} e^{-x} dx = c\Gamma(p_1 + p_2) = 1$$

или

$$c = \frac{1}{\Gamma(p_1 + p_2)}.$$

Плотность $f_\xi(x)$ имеет вид:

$$f_\xi(x) = \begin{cases} \frac{\lambda^{p_1+p_2}}{\Gamma(p_1+p_2)} x^{p_1+p_2-1} e^{-\lambda x}, & x > 0 \\ 0, & x \leqslant 0. \end{cases}$$

В ходе доказательства леммы 8 получено тождество, связывающее бета-функцию с гамма-функциями:

$$B(p_1, p_2) = \int_0^1 s^{p_1-1}(1-s)^{p_2-1} ds = \frac{\Gamma(p_1)\Gamma(p_2)}{\Gamma(p_1+p_2)}.$$

Следствие 2

Пусть случайные величины δ_i *,*$i = 1, \ldots, m$ *взаимно независимы, одинаково распределены и подчиняются стандартному нормальному распределению,* $\delta_i \sim N(0,1)$*. Тогда* $\delta_1^2 + \ldots + \delta_m^2 \sim G(\frac{1}{2}, \frac{m}{2})$.

Доказательство следует из лемм 7, 8.

Замечание 4

Доказано важное утверждение: распределение хи-квадрат с m степенями свободы является частным случаем гамма распределения с параметрами $p = 1/2$, $\lambda = m/2$.

Если случайная величина τ подчиняется распределению хи-квадрат с m степенями свободы, то ее плотность имеет вид:

$$f_\tau(x) = \begin{cases} \dfrac{x^{\frac{m}{2}-1} e^{-\frac{1}{2}x}}{2^{\frac{m}{2}} \Gamma(\frac{m}{2})}, & x > 0; \\ 0, & x \leq 0. \end{cases}$$