Критерии согласия и проверка гипотез об однородности выборок

А.Г. Трофимов

atrofimov@datalearning.ru
http://datalearning.ru

Курс “Математическая статистика”

Апрель 2015

Критерии согласия

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений случайной величины X

Статистическая гипотеза $H_0 : F_X(x, \theta) = G(x)$

$G(x)$ - заданное распределение

Вопрос:

Могло ли случиться так, что выборка $x_1, \ldots, x_n$ была получена из генеральной совокупности с распределением $G(x)$?

Определение

Критерии согласия (goodness of fit tests) - статистические критерии для проверки гипотез о виде распределения наблюдаемой генеральной совокупности

Общий подход к проверке критериев

Оценкой неизвестной функции распределения $F_X(x, \theta)$, рассчитанной по выборке $x_1, \ldots, x_n$, является *эмпирическая функция распределения* (ЭФР)

$F_n^*(x) = \frac{1}{n} \sum_{i=1}^{n} [x_i < x]$

$F_n^*(x)$ - реализация случайной ЭФР $\mathscr{F}_n^*(x)$

$\mathscr{F}_n^*(x)$ - состоятельная оценка функции распределения $F_X(x, \theta)$

Тогда в условиях гипотезы H_0:

$P[\Delta(F_n^*(x), G(x)) > \delta] \to 0$ при $n \to \infty, \forall \delta > 0$

$\Delta(F_n^*(x), G(x))$ - некоторая мера рассогласования между функциями $F_n^*(x)$ и $G(x)$

Критерии согласия:

- Критерий Колмогорова
- Критерий “омега-квадрат”
- Критерий Пирсона

Критерий Колмогорова

One-sample KS-test (А.Н.Колмогоров, Н.В.Смирнов, 1933)

Метрика Колмогорова:

$$\Delta(F_n^*(x), G(x)) = D_n = \sup_x |F_n^*(x) - G(x)|$$

Расчёт на практике:

$$D_n = \max_{i=\overline{1,n}} \left\{ \left| G(x_{(i)}) - \frac{2i-1}{2n} \right| + \frac{1}{2n} \right\}$$

Гипотеза:

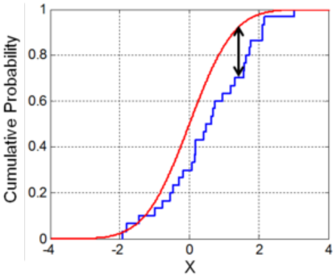
$$H_0 : F_X(x) = G(x)$$

Статистика критерия:

$$Z = \sqrt{n} D_n$$

$$Z|_{H_0} \sim KS$$

$$F_{KS}(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}$$



Критерий "омега-квадрат"

One-sample ω^2-test (H.Cramer, R. von Mises, 1930)

Метрика "омега-квадрат":

$$\Delta(F_n^*(x), G(x)) = \omega_n^2 = \frac{1}{n}\sum_{i=1}^{n} |F_n^*(x_i) - G(x_i)|^2$$

Расчёт на практике:

$$\omega_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(G(x_{(i)}) - \frac{2i-1}{2n}\right)^2$$

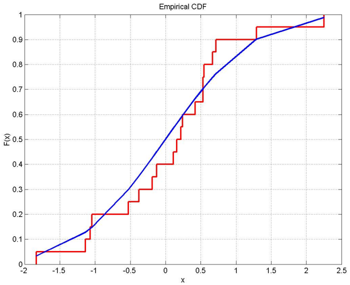
Гипотеза:

$$H_0 : F_X(x) = G(x)$$

Статистика критерия:

$$Z = n\omega_n^2$$

$$Z|_{H_0} \sim \omega^2$$



Критерий "хи-квадрат"

One-sample χ^2-test (K.Pearson, 1900)

Метрика "хи-квадрат":

$\chi_n^2 = \sum_{i=1}^{k} \frac{(\tilde{p}_i - p_i)^2}{p_i}$ k - число интервалов группировки

$p_i = \int\limits_{-\infty}^{\infty} g(x)dx,$ $\tilde{p}_i = \frac{n_i}{n}, i = \overline{1, k}$ - относительная частота

Гипотеза:

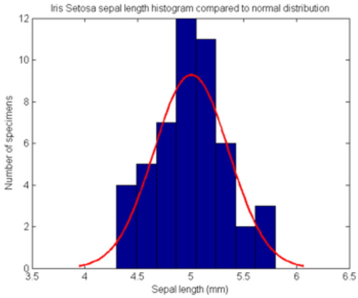
$H_0 : F_X(x) = G(x)$

Статистика критерия:

$Z = n\chi^2$

$Z|_{H_0} \sim \chi^2(k - r - 1)$

r - число оцененных параметров распределения



Критерий знаков

Определение

Выборки $x_1, \ldots, x_{n_X}$ и $y_1, \ldots, y_{n_Y}$ называются однородными, если они получены из одной генеральной совокупности

Определение

Выборки $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$ называются связанными (paired), если каждая пара (x_i, y_i) является наблюдением двумерного случайного вектора (X, Y)

Идея: если выборки получены из одной и той же генеральной совокупности, то значения x_i и y_i взаимозаменяемы

Критерий знаков

Критерий знаков (sign test):

$H_0: P[X-Y>0] = P[X-Y<0] = 0.5$

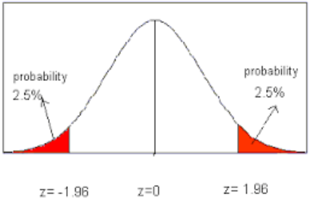
$K \sim N\left(np, \sqrt{np(1-p)}\right)$ - число положитильных разностей

$H = \frac{K}{n} \sim N\left(p, \sqrt{p(1-p)/n}\right)$ - частота положитильных разностей

Статистика критерия:

$Z = \frac{H-1/2}{\sqrt{1/4n}} = 2\sqrt{n}(H-1/2)$

$Z|_{H_0} \sim N(0,1)$



Критерий Манна-Уитни

Идея: если выборки $x_1, \ldots, x_{n_X}$ и $y_1, \ldots, y_{n_Y}$ получены из одной и той же генеральной совокупности, то элементы как первой, так и второй выборок в вариационном ряду объединённой выборки $x_1, \ldots, x_{n_X}, y_1, \ldots, y_{n_Y}$ перемешаны равномерно

Для оценки степени перемешивания данных двух выборок проводится ранжирование объединённой выборки

Определение

Рангом элемента z_i в выборке $z_1, \ldots, z_N$ называется его порядковый номер в вариационном ряду выборки

Пусть R_X – сумма рангов, соответствующих элементам из первой выборки

Критерий Манна-Уитни

Критерий Манна-Уитни (Mann-Whitney-Wilcoxon test, MWW-test, U-test, 1947):

Можно показать, что $\frac{n_X(n_X+1)}{2} \leq R_X \leq \frac{n_X(n_X+1)}{2} + n_X n_Y$

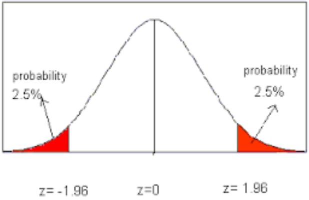
Введем статистику

$$U_X = n_X n_Y + \frac{n_X(n_X+1)}{2} - R_X, \quad 0 \leq U_X \leq n_X n_Y$$

Статистика критерия:

$$Z = \frac{U_X - \frac{n_X n_Y}{2}}{\sqrt{\frac{n_X n_Y (n_X + n_Y + 1)}{12}}}$$

$$Z|_{H_0} \sim N(0,1)$$



Двухвыборочный критерий Колмогорова

Two-sample KS-test (А.Н.Колмогоров, Н.В.Смирнов)

Метрика Колмогорова:

$$\Delta(F^*_{n_X}(x), G^*_{n_Y}(x)) = D_{n_X,n_Y} = \sup_x |F^*_{n_X}(x) - G^*_{n_Y}(x)|$$

Гипотеза:

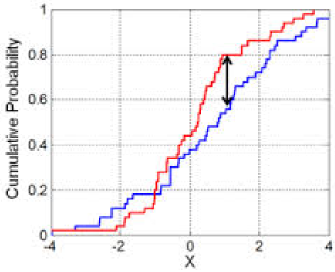
$$H_0 : F_X(x) = F_Y(x)$$

Статистика критерия:

$$Z = \sqrt{\frac{n_X n_Y}{n_X + n_Y}} D_{n_X,n_Y}$$

$$Z|_{H_0} \sim KS$$

$$F_{KS}(z) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 z^2}$$



Двухвыборочный критерий "омега-квадрат"

Two-sample ω^2-test (H.Cramer, R. von Mises)

Метрика "омега-квадрат":

$$\Delta(F^*_{n_X}(x), G^*_{n_Y}(x)) = \omega^2_{n_X,n_Y} = \frac{1}{n}\sum_{i=1}^{N} |F^*_{n_X}(z_i) - G^*_{n_Y}(z_i)|^2$$

$z_1, ..., z_N$ - объединенная выборка

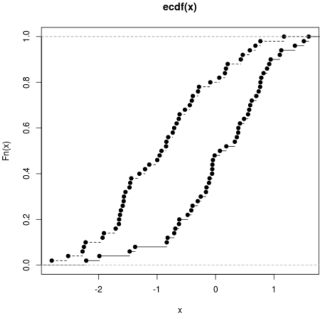
Гипотеза:

$H_0 : F_X(x) = F_Y(x)$

Статистика критерия:

$Z = \frac{n_X n_Y}{n_X + n_Y}\omega^2_{n_X,n_Y}$

$Z|_{H_0} \sim \omega^2$



Двухвыборочный критерий “хи-квадрат”

Two-sample χ^2-test (K.Pearson)

Метрика “хи-квадрат”:

$$\chi^2_{n_X,n_Y} = \sum_{i=1}^{k} \frac{1}{m_i^{(X)}+m_i^{(Y)}} \left(\frac{m_i^{(X)}}{n_X} - \frac{m_i^{(Y)}}{n_Y} \right)^2$$

k- число интервалов группировки

Гипотеза:

$H_0 : F_X(x) = F_Y(x)$

Статистика критерия:

$Z = n_x n_Y \chi^2_{n_X,n_Y}$

$Z|_{H_0} \sim \chi^2(k-1)$