Проверка статистических гипотез

А.Г. Трофимов

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

Курс "Математическая статистика"

Апрель 2015

Статистические гипотезы

Определение

Статистическая гипотеза (Statistical hypothesis) - любое предположение относительно параметров или закона распределения наблюдаемой случайной величины (или нескольких величин)

Основная гипотеза - H_0, альтернативная гипотеза - H'
Простые/сложные, параметрические/непараметрические

Примеры:

1. Случайная величина $X \sim B(1000, 0.06)$
2. Случайная величина $X \sim B(1000, p)$, где $0.04 \leq p \leq 0.08$
3. Дисперсия случайной величины X не более 2.3
4. Вероятность того, что во всей партии будет более 80 бракованных изделий, не превосходит 90

Статистический критерий

Дано:
Выборка $x_1, \dots, x_n$ наблюдений случайной величины X
Статистическая гипотеза H_0

Вопрос:
Могло ли случиться так, что выборка $x_1, \dots, x_n$ была получена из генеральной совокупности с указанными в гипотезе H_0 свойствами?

Определение

Статистический критерий (решающее правило) - правило, в соответствии с которым гипотеза H_0 принимается или отвергается

Статистика критерия

Определение

Статистика критерия (test statistic) - Статистика $Z = Z(X_1, ..., X_n)$, на основе реализации которой выдвигается *статистическое решение*. Реализация $z = Z(x_1, ..., x_n)$ статистики критерия, рассчитанная для выборки $x_1, \ldots, x_n$, называется выборочным значением статистики критерия

Свойства статистики критерия:

1. закон распределения $F_Z(z|H_0)$ известен
2. закон распределения $F_Z(z|H_0)$ чувствителен к факту справедливости основной или альтернативной гипотезы, т.е. законы распределения $F_Z(z|H_0)$ и $F_Z(z|H')$ должны существенно различаться

Критическая область

Допущение

Маловероятные события относительно статистики критерия Z считаются невозможными

Определение

Область допустимых значений статистики критерия Z (Region of acceptance) - область Ω_0 наиболее вероятных значений статистики критерия Z в условиях гипотезы H_0

Критическая область (Critical region) - область Ω' маловероятных значений статистики критерия Z в условиях гипотезы H_0

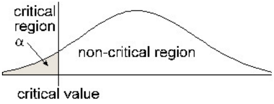
Статистический критерий

$z = Z(x_1, ..., x_n) \in \Omega_0 \Rightarrow H_0$ принимается

$z = Z(x_1, ..., x_n) \in \Omega' \Rightarrow H_0$ отклоняется

Типы критических областей

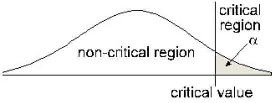
Left-tailed:



$$H_0 : m = m_0$$
$$H' : m < m_0$$

$$Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}$$
$$Z|_{H_0} \sim N(0;1)$$

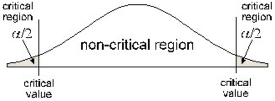
Right-tailed:



$$H_0 : m = m_0$$
$$H' : m > m_0$$

$$Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}$$
$$Z|_{H_0} \sim N(0;1)$$

Two-tailed:



$$H_0 : m = m_0$$
$$H' : m \neq m_0$$

$$Z = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}}$$
$$Z|_{H_0} \sim N(0;1)$$

Ошибки принятия статистического решения

Определение

Ошибка 1-го рода (Type I error) - ошибочное отклонение гипотезы H_0
Ошибка 2-го рода (Type II error) - ошибочное принятие гипотезы H_0

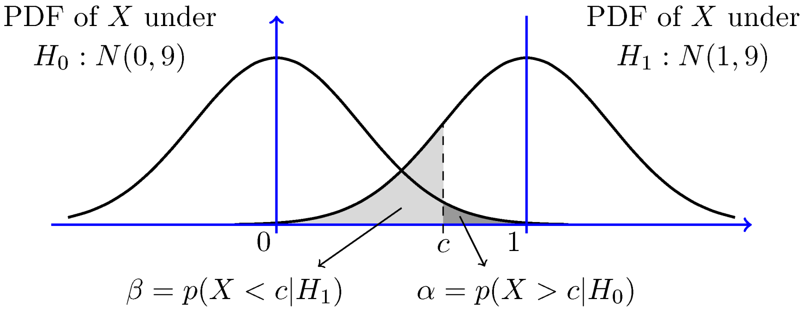
Стат. решение \ Факт	H_0 верна	H_0 не верна
H_0 принимается	правильное решение	ошибка 2-го рода
H_0 отвергается	ошибка 1-го рода	правильное решение

Пример:

H_0 : самолёт свой
H' : самолёт противника

Type I: сбит свой
Type II: пропущен чужой

Type I/II error tradeoff



Факт / Стат. решение	H_0 верна	H_0 не верна
H_0 принимается	$1-\alpha$	β
H_0 отвергается	α	$1-\beta$

α - уровень значимости (Significance level of a test)

$(1-\beta)$ - мощность критерия (Power of a test)

P-value

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений СВ $X \sim F_X(x, \theta)$

Определение

P-value - вероятность того, что статистика критерия в условиях гипотезы H_0 примет менее вероятные значения, чем она приняла для данной выборки

Формально:

$p = P[Z \leq z|_{H_0}]$ for left-tail test
$p = P[Z \geq z|_{H_0}]$ for right-tail test
$p = 2\min\{P[Z \leq z|_{H_0}], P[Z \geq z|_{H_0}]\}$ for two-tail test
$z = Z(x_1, ..., x_n)$

P-value

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений СВ $X \sim F_X(x, \theta)$

Определение

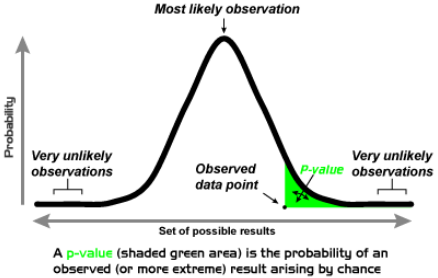
P-value - вероятность того, что статистика критерия в условиях гипотезы H_0 примет менее вероятные значения, чем она приняла для данной выборки

Формально:

$p = P[Z \leq z|_{H_0}]$ for left-tail test
$p = P[Z \geq z|_{H_0}]$ for right-tail test
$p = 2\min\{P[Z \leq z|_{H_0}], P[Z \geq z|_{H_0}]\}$ for two-tail test
$z = Z(x_1, ..., x_n)$

$\alpha = P[Z \leq z_1|_{H_0}]$ for left-tail test
$\alpha = P[Z \geq z_2|_{H_0}]$ for right-tail test
$\alpha = 2\min\{P[Z \leq z_1|_{H_0}], P[Z \geq z_2|_{H_0}]\}$ for two-tail test

P-value



P-value - наибольший уровень значимости, при котором гипотеза H_0 принимается

Статистический критерий

$p > \alpha \Rightarrow H_0$ принимается
$p < \alpha \Rightarrow H_0$ отклоняется

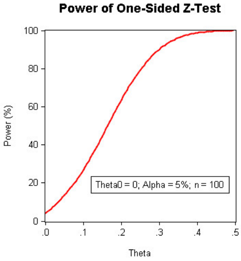
Функция мощности критерия

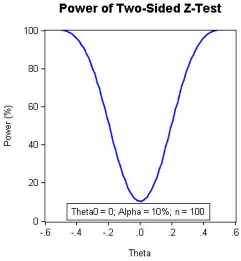
Дано: $H_0 : \theta = \theta_0,$ уровень значимости α

Определение

Функция мощности критерия - зависимость мощности критерия $(1 - \beta(\theta))$ от неизвестного параметра θ

Для каждой альтернативной гипотезы $H' : \theta = \theta_1$ расчитываем $\beta(\theta_1) = P[Z \in \Omega_0|_{H'}]$





Резюме

- Статистическая гипотеза (Statistical hypothesis)
- Простая (Simple) и сложная (Composite) гипотезы
- Основная (Null) и альтернативная (Alternative) гипотезы
- Статистический критерий (Statistical criterion)
- Статистика критерия (Test statistic)
- Область допустимых значений (Region of acceptance)
- Критическая область (Critical region)
- Критические точки (Critical points)
- Статистическое решение (Statistical decision)
- Ошибки 1-го и 2-го рода (Type I/II errors)
- Уровень значимости (Significance level)
- P-value
- Мощность критерия (Power of a test)

One-sample z-test

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений СВ $X \sim N(m, \sigma)$

σ - известно

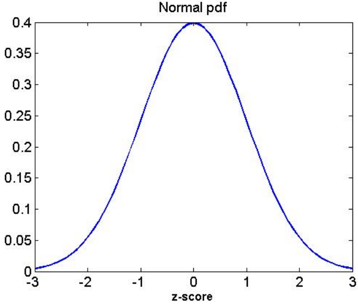
Гипотеза:

$H_0 : m = m_0$

Статистика критерия:

$$Z = \frac{\overline{X} - m_0}{\sigma / \sqrt{n}}$$

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



One-sample t-test

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений СВ $X \sim N(m, \sigma)$

σ - неизвестно

Гипотеза:

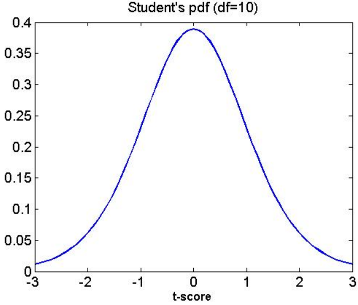
$H_0 : m = m_0$

Статистика критерия:

$$Z = \frac{\overline{X} - m_0}{S/\sqrt{n}}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$$Z|_{H_0} \sim T(n-1)$$



Chi-square variance test

Дано:

Выборка $x_1, \ldots, x_n$ наблюдений СВ $X \sim N(m, \sigma)$

Гипотеза:

$H_0 : \sigma = \sigma_0$

Статистика критерия:

m - известно:

$Z = \frac{nS_0^2}{\sigma_0^2}$

$S_0^2 = \frac{1}{n} \sum\limits_{i=1}^{n} (X_i - m)^2$

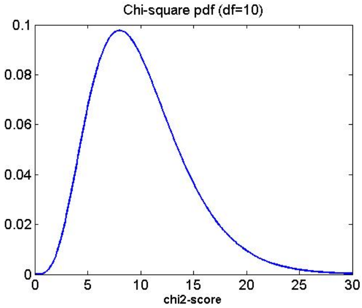
$Z|_{H_0} \sim \chi^2(n)$

m - неизвестно:

$Z = \frac{(n-1)S^2}{\sigma_0^2}$

$S^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} \left(X_i - \overline{X}\right)^2$

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



Two-sample z-test

Дано:

Выборка $x_1, \ldots, x_{n_1}$ наблюдений СВ $X \sim N(m_1, \sigma_1)$
Выборка $y_1, \ldots, y_{n_2}$ наблюдений СВ $Y \sim N(m_2, \sigma_2)$
σ_1, σ_2 - известны

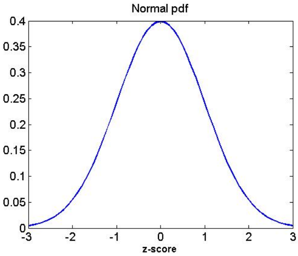
Гипотеза:

$H_0 : m_1 = m_2$

Статистика критерия:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

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



Two-sample t-test

Дано:

Выборка $x_1, \ldots, x_{n_1}$ наблюдений СВ $X \sim N(m_1, \sigma_1)$
Выборка $y_1, \ldots, y_{n_2}$ наблюдений СВ $Y \sim N(m_2, \sigma_2)$
σ_1, σ_2 - неизвестны, но $\sigma_1 = \sigma_2$

Гипотеза:

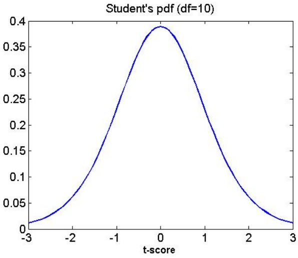
$H_0 : m_1 = m_2$

Статистика критерия:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{S/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

$$Z|_{H_0} \sim T(n_1 + n_2 - 1)$$



Welch's t-test

Дано:

Выборка $x_1, \ldots, x_{n_1}$ наблюдений СВ $X \sim N(m_1, \sigma_1)$

Выборка $y_1, \ldots, y_{n_2}$ наблюдений СВ $Y \sim N(m_2, \sigma_2)$

σ_1, σ_2 - неизвестны, но $\sigma_1 \neq \sigma_2$

Гипотеза:

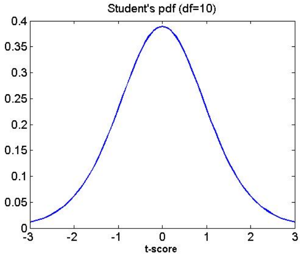
$H_0 : m_1 = m_2$

Статистика критерия:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$Z|_{H_0} \sim T([1/k])$$

$$k = \frac{\left(\frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2/n_2}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_2 - 1}$$



Two-sample F-test

Дано:

Выборка $x_1, \ldots, x_{n_1}$ наблюдений СВ $X \sim N(m_1, \sigma_1)$
Выборка $y_1, \ldots, y_{n_2}$ наблюдений СВ $Y \sim N(m_2, \sigma_2)$

Гипотеза:

$H_0 : \sigma_1 = \sigma_2$

Статистика критерия:

m_1, m_2 - известны:

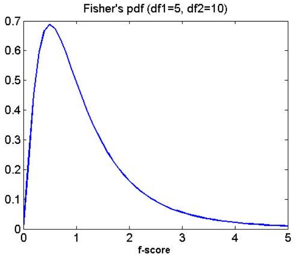
$Z = \frac{S_{01}^2}{S_{02}^2}$

$Z|_{H_0} \sim F(n_1, n_2)$

m_1, m_2 - неизвестны:

$Z = \frac{S_1^2}{S_2^2}$

$Z|_{H_0} \sim F(n_1 - 1, n_2 - 1)$



Резюме

Hypothesis H_0	Expectation	Variance	Statistical test
$m = m_0$	*unknown*	σ	one-sample z-test
$m = m_0$	*unknown*	*unknown*	one-sample t-test
$\sigma = \sigma_0$	*known/unknown*	*unknown*	chi-square variance test
$m_1 = m_2$	*unknown*	*unknown*, $\sigma_1 = \sigma_2$	two-sample t-test
$m_1 = m_2$	*unknown*	*unknown*, $\sigma_1 \neq \sigma_2$	Welch's t-test
$\sigma_1 = \sigma_2$	*known/unknown*	*unknown*	two-sample F-test

Алгоритм проверки статистических гипотез

1. Сформулировать проверяемую гипотезу H_0 и альтернативную гипотезу H'
2. Выбрать уровень значимости α
3. Выбрать статистику критерия Z
4. Найти закон распределения статистики Z при условии истинности основной гипотезы H_0
5. Построить область допустимых значений Ω_0 и критическую область Ω'
6. Вычислить выборочное значение статистики критерия $z = Z(x_1, ..., x_n)$ или *p-value*
7. Принять статистическое решение

One-proportion z-test

Пусть проводится серия из n испытаний в схеме Бернулли
Случайная величина K – число "успехов"

One-proportion z-test

Пусть проводится серия из n испытаний в схеме Бернулли
Случайная величина K – число "успехов"
$K \sim B(n,p) \quad \mathrm{M}[K] = np \quad \mathrm{D}[K] = np(1-p)$
Аппроксимация $K \sim N(np, \sqrt{np(1-p)})$
Частота "успеха" $H = K/n \sim N(p, \sqrt{p(1-p)/n})$

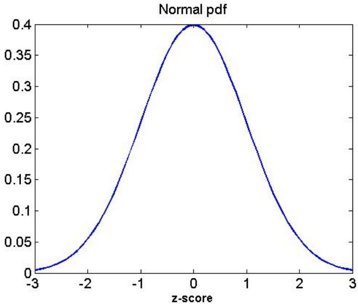
Гипотеза:

$H_0 : p = p_0$

Статистика критерия:

$$Z = \frac{H - p_0}{\sqrt{p_0(1-p_0)/n}}$$

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



Two-proportion z-test

Пусть проводится две серии испытаний в схеме Бернулли
Частота "успеха" в 1-ой серии
$H_1 = K_1/n \sim N(p_1, \sqrt{p_1(1-p_1)/n_1})$
Частота "успеха" во 2-ой серии
$H_2 = K_2/n \sim N(p_2, \sqrt{p_2(1-p_2)/n_2})$

Гипотеза:

$H_0 : p_1 = p_2$

Статистика критерия:

$$Z = \frac{H_1 - H_2}{\sqrt{H(1-H)}\sqrt{1/n_1 + 1/n_2}}$$
$$H = \frac{n_1 H_1 + n_2 H_2}{n_1 + n_2}$$

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