МС-20-21 Теоретический материал

Основные понятия статистической проверки гипотез

Пусть $\vec{X} = (X_1, \ldots, X_n)$ — случайная выборка объема n из некоторого генерального распределения. Не ограничивая общности можно считать, что существует определенная схема испытаний, при осуществлении которой вычисляется случайная величина X, а $X_1, \ldots, X_n$ — это те ее значения, которые X принимает в результате серии n независимых испытаний. Таким образом, случайные величины $X_1, \ldots, X_n$ независимы и распределены по тому же закону, что и X.

Статистической гипотезой называется любое утверждение о виде или параметрах генерального распределения.

Статистическая гипотеза называется **параметрической**, если она основана на предположении, что генеральное распределение известно с точностью до конечного числа параметров.

Рассмотрим базисное предположение, состоящее в том, что генеральное распределение зависит от некоторого параметра $\theta \in \mathbb{R}^n$. Параметрическая гипотеза называется **простой**, если она имеет вид: $\theta = \theta_0$, где θ_0 – некоторое фиксированное значение параметра θ. Гипотеза вида: $\theta \in \Theta$, где Θ – какое-либо множество, содержащее, по меньшей мере, два различных элемента, называется **сложной**.

Пусть H_0 и H_1 — две взаимоисключающие статистические гипотезы.

Проверяемая гипотеза H_0 называется **основной**, а дополнительная гипотеза H_1 — **альтернативной**. Предполагается, что одна из этих гипотез выполняется.

Статистическим критерием с критической областью $K \subset \mathbb{R}^n$ называется правило, в соответствии с которым H_0 отвергается, если выборка попадает в критическую область, $(X_1, \ldots, X_n) \in K$.

Критические области задаются либо при помощи неравенств вида $K = \{t < c_1\}$ или $K = \{t > c_2\}$, либо как объединение $K = \{t < c_1\} \cup \{t > c_2\}$, где $t = t(x_1, \ldots, x_n)$ — подходящая функция от выборочных значений, а c_1 и c_2 — некоторые константы, такие что $c_1 < c_2$.

Во всех этих случаях числа c_1 и c_2 называются **критическими значениями**, а функция $t(x_1, \ldots, x_n)$ — **статистикой критерия**. **Статистикой критерия** называется также случайная величина $T = t(X_1, \ldots, X_n)$.

Ошибка первого рода состоит в том, что отвергается верная гипотеза H_0 . **Ошибка второго рода** состоит в том, что отвергается верная гипотеза H_1 .

Вероятность ошибки первого рода называется **уровнем значимости критерия** и обозначается α.

Вероятность ошибки второго рода обозначается β, а величина 1 − β называется **мощностью критерия**.

ошибка I рода	***ошибка II рода***
Отвергается основная (нулевая) гипотеза, хотя она верна.	Отвергается конкурирующая гипотеза, хотя она верна.
Вероятность ошибки $P(H_1\|H_0) = \alpha$, α – ***уровень значимости критерия*** (обычно $\alpha = 0{,}05$; 0,01; 0,005; 0,001).	Вероятность ошибки $P(H_0\|H_1) = \beta$ (величина β, как правило, заранее неизвестна)
Вероятность принять верную (нулевую) гипотезу $P(H_0\|H_0) = 1 - \alpha$.	Вероятность принять верную (конкурирующую) гипотезу $P(H_1\|H_1) = 1 - \beta$, $(1-\beta)$ – ***мощность критерия***.

Пусть $T(\vec{X})$ — некоторая статистика, характеризующая отклонение эмпирических данных от тех гипотетических значений, которые соответствуют проверяемой гипотезе H_0. И пусть, кроме того, распределение этой статистики, в случае справедливости гипотезы H_0, известно (точно или хотя бы приближенно).

Обозначим $T_{\text{набл}}$ — **наблюдаемое (или выборочное)** значение статистики, т. е. значение статистики $T(\vec{x})$, вычисленное для полученной реализации случайной выборки $\vec{x} = (x_1, \ldots, x_n)$. Зафиксируем достаточно малое число $\alpha \in (0; 1)$. Разобьем множество всех возможных значений статистики T на две части: критическую область критерия K и его дополнение $\overline{K} = D$. Критическая область K выбирается так, чтобы выполнялось соотношение

$$P_{H_0}(\{T \in K\}) \leq \alpha.$$

(Через $P_{H_0}(A)$ обозначена вероятность события A, вычисленная в предположении, что гипотеза H_0 верна.) Таким образом, критическая область включает в себя маловероятные значения статистики Т, при условии, что верна основная гипотеза H_0.

Критерий проверки гипотезы теперь можно сформулировать следующим образом:
По имеющейся выборке $x_1, \ldots, x_n$ находится наблюдаемое значение статистики $T_{\text{набл}}$.
Если окажется, что $T_{\text{набл}} \in K$, то делается вывод о том, что в предположении справедливости гипотезы H_0 произошло маловероятное событие. Поэтому эта гипотеза должна быть отвергнута, как противоречащая статистическим данным $x_1, \ldots, x_n$, полученным в результате эксперимента. В противном случае (т. е. если $T_{\text{набл}} \in D = \overline{K}$) считается, что данные не противоречат H_0.

Замечание. Если $T_{\text{набл}} \notin K$, то гипотеза H_0 принимается, но сам по себе тот факт, что $T_{\text{набл}} \notin K$, не является доказательством истинности H_0. Также и тот факт, что $T_{\text{набл}} \in K$, не является доказательством истинности H_1.

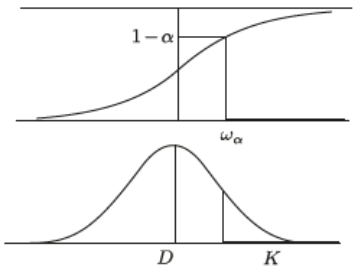
Способ выбора критической области K зависит от вида статистики T и гипотез H_0 и H_1.

Как правило, критическая область задается одним из следующих трех способов.

Правосторонняя критическая область:

$K = \{\vec{x}: T(\vec{x}) > c\}$

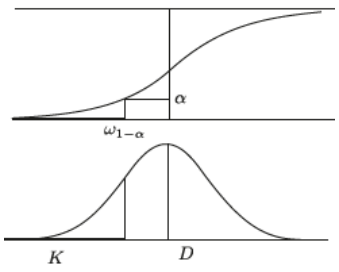
Гипотеза H_0 отклоняется, если $T_{\text{набл}} > \omega_\alpha$



Левосторонняя критическая область:

$K = \{\vec{x}: T(\vec{x}) < c\}$

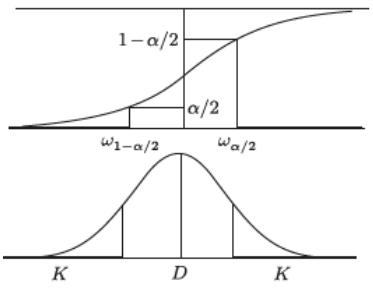
Гипотеза H_0 отклоняется, если $T_{\text{набл}} < \omega_{1-\alpha}$



Двусторонняя критическая область:

$K = \{\vec{x}: T(\vec{x}) < c_1\} \cup \{\vec{x}: T(\vec{x}) > c_2\}.$

Гипотеза H_0 отклоняется, если $T_{\text{набл}} \in (-\infty; \omega_{1-\frac{\alpha}{2}}) \cup (\omega_{\frac{\alpha}{2}}; +\infty)$



ω_q — $100q$%-я точка статистки критерия, α — уровень значимости.

Лемма Неймана–Пирсона

Предположим, что генеральное распределение имеет зависящую от параметра θ положительную при всех x плотность $f(x;\theta) > 0$. Пусть H_0 и H_1 – простые гипотезы вида $H_0 : \theta = \theta_0$ и $H_1 : \theta = \theta_1$.

Запишем функции правдоподобия, соответствующие этим гипотезам:

$$L_0(\theta_0, x_1 \dots, x_n) = f(x_1;\theta_0) \cdot \dots \cdot f(x_n;\theta_0),$$
$$L_1(\theta_1, x_1 \dots, x_n) = f(x_1;\theta_1) \cdot \dots \cdot f(x_n;\theta_1)$$

Теорема (лемма Неймана–Пирсона). Для любого $\alpha \in (0,1)$ существует такая константа c_α, что критерий с критической областью

$$\frac{L_1(\theta_1, x_1 \dots, x_n)}{L_0(\theta_0, x_1 \dots, x_n)} > c_\alpha,$$

является наиболее мощным критерием среди всех статистических критериев с какой-либо критической областью K, предназначенных для проверки H_0 против H_1 с уровнем значимости α.

P-значение критерия

Пусть имеется статистический критерий $T(\vec{X})$ с критической областью K. И пусть получена реализация $\vec{x}$ случайной выборки $\vec{X} = (X_1, \dots, X_n)$.

***P*-значением (*P*-value)** $p(\vec{x})$ статистического критерия $T(\vec{x})$ называется наименьшая величина уровня значимости, при котором нулевая гипотеза отклоняется:

$$p(\vec{x}) = \min\{\alpha : T(\vec{x}) \in K\}.$$

Величина $p(\vec{x})$ задает фактический уровень значимости. Для всех значений уровня значимости, таких, что $\alpha \leq p(\vec{x})$, гипотеза H_0 принимается, при всех $\alpha > p(\vec{x})$ гипотеза H_0 отклоняется.

Чем меньше P-значение, тем сильнее основания отклонить нулевую гипотезу.

Теорема 1. Пусть $T = T(\vec{X})$ — статистика критерия, $T_{\text{набл}} = T(\vec{x})$ — наблюдаемое значение статистики критерия.

Если критическая область имеет вид: $\{T_{\text{набл}} > \omega_\alpha\}$, где ω_α — 100α-процентная точка статистики T (**правосторонняя критическая область**), то P-значение $p_1(\vec{x})$ находится по формуле:

$$p_1(\vec{x}) = P_{H_0}(\{T > T_{\text{набл}}\}),$$

где через $P_{H_0}(A)$ обозначена вероятность события A, вычисленная в предположении справедливости гипотезы H_0.

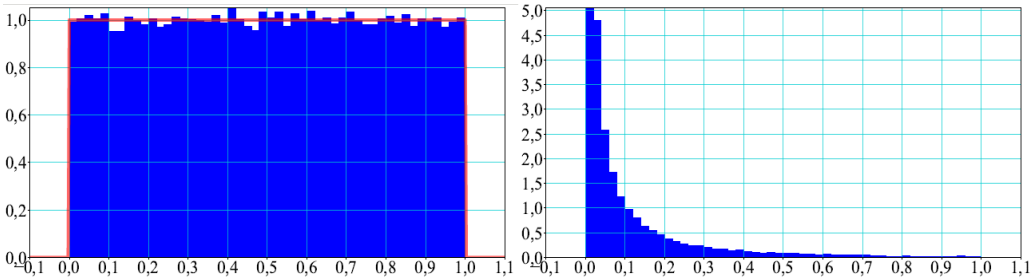
Если критическая область имеет вид: $\{T_{\text{набл}} < \omega_{1-\alpha}\}$ (**левосторонняя критическая область**), то P-значение $p_2(\vec{x})$ находится по формуле:

$$p_2(\vec{x}) = P_{H_0}(\{T < T_{\text{набл}}\}) = 1 - p_1(\vec{x}).$$

Если критическая область имеет вид: $T_{\text{набл}} \in (-\infty; \omega_{1-\frac{\alpha}{2}}) \cup (\omega_{\frac{\alpha}{2}}; +\infty)$ (**двусторонняя критическая область**), то P-значение $p_3(\vec{x})$ находится по формуле:

$$p_3(\vec{x}) = 2 \cdot \min\{p_1(\vec{x}), p_2(\vec{x})\}.$$

Теорема 2. Если H_0 – верна, то случайная величина $PV \sim Unif[0;1]$ (то есть распределение с.в. PV является равномерным на $[0;1]$).



Распределение с.в. PV , когда

1) H_0 – верна; 2) H_0 – отвергается.

Проверка гипотезы об определенном значении параметра нормального распределения

1) Проверка гипотезы об определенном значении генерального среднего при <u>известной дисперсии</u> на уровне значимости α.

$H_0: \mu = \mu_0$

против любой из трех альтернативных гипотез H_1: 1) $\mu > \mu_0$; 2) $\mu < \mu_0$; 3) $\mu \neq \mu_0$.

Статистика $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$.

H_1	K
$\mu > \mu_0$	$Z > z_{1-\alpha} = -z_\alpha$
$\mu < \mu_0$	$Z < z_\alpha = -z_{1-\alpha}$
$\mu \neq \mu_0$	$\lvert Z\rvert > z_{1-\alpha/2}$

$z_{1-\alpha/2}$ — квантиль стандартного нормального распределения уровня $1-\alpha/2$.

2) Проверка гипотезы об определенном значении генерального среднего при <u>неизвестной дисперсии</u> на уровне значимости α.

$H_0: \mu = \mu_0$; H_1: 1) $\mu > \mu_0$; 2) $\mu < \mu_0$; 3) $\mu \neq \mu_0$.

Статистика $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$.

H_1	K
$\mu > \mu_0$	$t > t_{1-\alpha}(n-1)$
$\mu < \mu_0$	$t < t_\alpha(n-1) = -t_{1-\alpha}(n-1)$
$\mu \neq \mu_0$	$\lvert t\rvert > t_{1-\frac{\alpha}{2}}(n-1)$

$t_{1-\alpha/2}\,(n-1)$ – квантиль распределения Стьюдента с $n-1$ степенями свободы уровня $1-\frac{\alpha}{2}$.

3) Проверка гипотезы об определенном значении генеральной дисперсии при <u>известном генеральном среднем</u> μ на уровне значимости α.

$H_0: \sigma^2 = \sigma_0^2$; H_1: 1) $\sigma^2 > \sigma_0^2$; 2) $\sigma^2 < \sigma_0^2$; 3) $\sigma^2 \neq \sigma_0^2$.

Статистика $\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{ns_0^2}{\sigma_0^2} \sim \chi^2(n)$.

H_1	K
$\sigma^2 > \sigma_0^2$	$\chi^2 > \chi^2{}_{1-\alpha}(n)$
$\sigma^2 < \sigma_0^2$	$\chi^2 < \chi^2{}_{\alpha}(n) = -\chi^2{}_{1-\alpha}(n)$
$\sigma^2 \neq \sigma_0^2$	$\{\chi^2 < \chi^2{}_{\frac{\alpha}{2}}(n)\} \cup \{\chi^2 > \chi^2{}_{1-\frac{\alpha}{2}}(n)\}$

$\chi^2_\alpha\,(n)$ – квантиль распределения χ^2 с n степенями свободы уровня α.

4) Проверка гипотезы об определенном значении генеральной дисперсии при неизвестном генеральном среднем μ на уровне значимости α.

$H_0: \sigma^2 = \sigma_0^2;\ H_1: 1)\ \sigma^2 > \sigma_0^2; 2)\ \sigma^2 < \sigma_0^2; 3)\ \sigma^2 \neq \sigma_0^2.$

Статистика $\chi^2 = \frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{(n-1)s^2}{\sigma_0^2} \sim \chi^2(n-1).$

Критическая область определяется той же таблицей, что и в п.3), но с числом степеней свободы $m = n - 1.$

5) Гипотеза о неизвестной вероятности успеха в испытаниях Бернулли на уровне значимости α

$H_0: p = p_0\ ;\ H_1: 1)\ p > p_0; 2)\ p < p_0; 3)\ p \neq p_0.$

Статистика $Z = \frac{w - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$, где $w = m/n$ – относительная частота успехов в n наблюдениях

Далее критические точки и области для проверки выбираются так же, как при проверке гипотезы о неизвестном среднем при известной дисперсии.

Замечание. Этим методом можно пользоваться только при больших объемах выборки (порядка нескольких десятков или сотен).

Python

1) **ztest** - Тест для среднего при известной дисперсии

https://www.statsmodels.org/dev/generated/statsmodels.stats.weightstats.ztest.html

2) Для проверки гипотезы **о числовом значении математического ожидания** используется функция **ttest_1sample(*x*, popmean)** модуля **scipy.stats**.

Параметры:

x – выборка,

popmean – гипотетическое значение математического ожидания.

Функция возвращает наблюдаемое значение статистики и *p*-значение критерия.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mstats.ttest_1samp.html

3) Проверяется гипотеза **о вероятности p некоторого события A**, $H_0: p=p_0$ против альтернатив вида $H_1: p>p_0$, $H_2: p<p_0$, $H_3: p\neq p_0$ на уровне значимости α.

Функция

binom_test(x, n, p0, alternative)

модуля **scipy.stats** использует точный биномиальный тест, используя статистику $T=x$ (x – число успехов).

Параметры:

x – число успехов,

n – число испытаний,

p0 – гипотетическое значение вероятности,

alternative – вид конкурирующей гипотезы (“two-sided”, “greater”, “less”).

(Второй вариант задания данных: x – набор двух значений, числа успехов и числа неудач. В этом случае параметр n игнорируется).

Возвращает p-значение критерия.