МС-23 Теоретический материал
Сравнение генеральных средних и генеральных дисперсий двух нормальных совокупностей, проверка гипотезы о равенстве вероятностей двух событий

Пусть $\vec{X} = (X_1, \ldots, X_m)$ — выборка из $N(\mu_x, \sigma_x^2)$, $\vec{Y} = (Y_1, \ldots, Y_n)$ — выборка из $N(\mu_y, \sigma_y^2)$. Далее считаем, что выборки $\vec{X}$ и $\vec{Y}$ независимыми, что означает независимость в совокупности $m+n$ случайных величин $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

1) Сравнение генеральных средних при известной дисперсии
(σ_x^2, σ_y^2 — известны, μ_x, μ_y – неизвестны)

$H_0 : \mu_x = \mu_y$
против любой из трех альтернативных гипотез H_1 : 1) $\mu_x > \mu_y$; 2) $\mu_x < \mu_y$; 3) $\mu_x \neq \mu_y$.

Статистика $Z = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}$.

H_1	K
1) $\mu_x > \mu_y$	$Z > Z_\alpha$
2) $\mu_x < \mu_y$	$Z < -Z_\alpha$
3) $\mu_x \neq \mu_y$	$\lvert Z \rvert > Z_{\alpha/2}$

Z_α — процентная точка стандартного нормального распределения $N(0,1)$.

2) Сравнение генеральных средних при неизвестных и равных дисперсиях
($\sigma_x^2 = \sigma_y^2 = \sigma$ – неизвестны)

$H_0 : \mu_x = \mu_y$; H_1 : 1) $\mu_x > \mu_y$; 2) $\mu_x < \mu_y$; 3) $\mu_x \neq \mu_y$.

Статистика $T = \frac{\bar{X}-\bar{Y}}{s\sqrt{\frac{1}{m}+\frac{1}{n}}}$,

$s^2 = \frac{m-1}{m+n-2}s_x^2 + \frac{n-1}{m+n-2}s_y^2$

H_1	K
1) $\mu_x > \mu_y$	$T > t_\alpha(m+n-2)$
2) $\mu_x < \mu_y$	$T < -t_\alpha(m+n-2)$
3) $\mu_x \neq \mu_y$	$\lvert T \rvert > t_{\alpha/2}(m+n-2)$

$t_\alpha\,(m+n-2)$ – 100α -процентная точка распределения Стьюдента с $m+n-2$ степенями свободы

3) Сравнение генеральных средних с неизвестными и неравными дисперсиями
Задача сравнения средних двух нормально распределенных совокупностей при неизвестных и неравных дисперсиях известна как проблема **Беренса-Фишера**. Точного решения этой задачи до настоящего времени нет.

Одно из приближений даёт критерий **Кохрана–Кокса**.

Статистика критерия:

$$T = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}};$$

Критическое значение статистики:

$$t_\alpha = \frac{\frac{s_x^2}{m}t_{\alpha;m-1}+\frac{s_y^2}{n}t_{\alpha;n-1}}{\frac{s_x^2}{m}+\frac{s_y^2}{n}}.$$

Если выполняется неравенство $|T_{\text{набл.}}| > t_\alpha$, гипотеза H_0 отклоняется.

Структура критерия по проверке гипотезы **о равенстве дисперсий двух нормальных распределений** зависит от того, известно или нет генеральное среднее, а также от вида альтернативной гипотезы.

4) Сравнение дисперсий двух нормальных распределений
($\mu_x, \mu_y, \sigma_x^2, \sigma_y^2$ – неизвестны)

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

Статистика $F = \frac{S_x^2}{S_y^2}$.

H_1	Критическая область
1) $\sigma_x^2 > \sigma_y^2$	$\frac{s_x^2}{s_y^2} > F_\alpha(m-1, n-1)$
2) $\sigma_x^2 < \sigma_y^2$	$\frac{s_y^2}{s_x^2} > F_\alpha(n-1, m-1)$
3) $\sigma_x^2 \neq \sigma_y^2$	$\frac{s_1^2}{s_2^2} > F_{\alpha/2}(k_1, k_2)$

символ	$s_x^2 \geq s_y^2$	$s_x^2 < s_y^2$
s_1^2	s_x^2	s_y^2
s_2^2	s_y^2	s_x^2
k_1	$m-1$	$n-1$
k_2	$n-1$	$m-1$

$F_\alpha(k_1, k_2)$ – 100α -процентная точка распределения Фишера с k_1 и k_2 степенями свободы.

5) Сравнение дисперсий двух нормальных распределений
(μ_x, μ_y – известны; σ_x^2, σ_y^2 – неизвестны)

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

Статистика $F = \frac{S_{0x}^2}{S_{0y}^2}$.

$$S_{0x}^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \mu_x)^2;\ S_{0y}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \mu_y)^2;$$

$F_\alpha(k_1, k_2)$ – 100α -процентная точка распределения Фишера с k_1 и k_2 степенями свободы, а критическая область определяется той же таблицей, что и в п.4), но с числом степеней свободы на единицу больше.

6) Гипотеза о равенстве вероятностей успеха в двух сериях испытаний Бернулли (n_1 и n_2 порядка сотен или более**).**
$H_0: p_1 = p_2$; H_1: 1) $p_1 > p_2$; 2) $p_1 < p_2$; 3) $p_1 \neq p_2$.

Статистика $Z = \frac{w_1 - w_2}{\sqrt{w(1-w)(\frac{1}{n_1}+\frac{1}{n_2})}}$, где $w_1 = \frac{m_1}{n_1}$, $w_2 = \frac{m_2}{n_2}$, а $w = \frac{m_1+m_2}{n_1+n_2}$ - относительная частота успехов в объединенных сериях испытаний.
Далее критические точки и области для проверки выбираются так же, как при сравнении генеральных средних при известной дисперсии.

Python

scipy.stats.ttest_ind

двусторонний тест для нулевой гипотезы о том, что две независимые выборки имеют идентичные средние (ожидаемые) значения.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html#scipy.stats.ttest_ind

https://www.kite.com/python/docs/statsmodels.stats.weightstats.DescrStatsW