Сводная таблица по математической статистике

Python для всех пунктов:

1. Двусторонний тест: p-value $= 2 \cdot \min\left\{\text{r.cdf}(v_p), 1 - \text{r.cdf}(v_p)\right\}$, где r – распределение статистики, v_p – значение расчётной статистики.
2. $\overline{X} =$ np.mean(x)
3. $S_0 =$ np.std$(x, ddof = 1)$, $S_0^2 =$ np.var$(x, ddof = 1)$

1. Одно распределение

Название	Предпосылки	H_0	H_1	Статистика	Выводы	Python (numpy, scipy.stats)
Гипотеза о матожидании	1. $X \sim N(\mu, \sigma^2)$ 2. σ^2 - известно	$\mu = \mu_0$	$\mu \neq \mu_0$	$z_p = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	Не отвергаем на уровне значимости α, если 1. $z_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right)$ 3. p-value $> \alpha$	1. $z_{1-\frac{\alpha}{2}} =$ norm.ppf$(q = 1 - \alpha/2)$, 2. p-value $= 2 \cdot \left(1 - \text{norm.cdf}(\text{abs}(z_p))\right)$
Гипотеза о матожидании	1. $X \sim N(\mu, \sigma^2)$ 2. σ^2 - неизвестно	$\mu = \mu_0$	$\mu \neq \mu_0$	$t_p = t^{(n-1)} = \dfrac{\overline{X} - \mu_0}{S_0/\sqrt{n}} \sim T_{n-1}$	Не отвергаем на уровне значимости α, если 1. $t_p \in \left(-t^{(n-1)}_{1-\frac{\alpha}{2}}, t^{(n-1)}_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{X} - t^{(n-1)}_{1-\frac{\alpha}{2}}\dfrac{S_0}{\sqrt{n}}, \overline{X} + t^{(n-1)}_{1-\frac{\alpha}{2}}\dfrac{S_0}{\sqrt{n}}\right)$ 3. p-value $> \alpha$	1. $t^{(n-1)}_{1-\frac{\alpha}{2}} =$ t.ppf$(df = n-1, q = 1 - \alpha/2)$, 2. p-value $= 2 \cdot \left(1 - \text{t.cdf}(\text{abs}(t_p), df = n-1)\right)$
Гипотеза о дисперсии	1. $X \sim N(\mu, \sigma^2)$ 2. μ - неизвестно	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$C_p = C^{(n-1)} = \dfrac{S_0^2(n-1)}{\sigma_0^2} \sim \chi^2_{n-1}$	Не отвергаем на уровне значимости α, если 1. $C_p \in \left(C^{(n-1)}_{\frac{\alpha}{2}}, C^{(n-1)}_{1-\frac{\alpha}{2}}\right)$, 2. $\sigma_0^2 \in \left(\dfrac{(n-1)S_0^2}{C^{(n-1)}_{1-\frac{\alpha}{2}}}, \dfrac{(n-1)S_0^2}{C^{(n-1)}_{\frac{\alpha}{2}}}\right)$ 3. p-value $> \alpha$	1. $C^{(n-1)}_{\frac{\alpha}{2}} =$ chi2.ppf$(df = n-1, q = \alpha/2)$, 2. $C^{(n-1)}_{1-\frac{\alpha}{2}} =$ chi2.ppf$(df = n-1, q = 1 - \alpha/2)$, 3. p-value
Гипотеза о дисперсии	1. $X \sim N(\mu, \sigma^2)$ 2. μ - известно	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$C_p = C^{(n)} = \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{\sigma_0^2} \sim \chi^2_n$	Не отвергаем на уровне значимости α, если 1. $C_p \in \left(C^{(n)}_{\frac{\alpha}{2}}, C^{(n)}_{1-\frac{\alpha}{2}}\right)$, 2. $\sigma_0^2 \in \left(\dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{C^{(n-1)}_{1-\frac{\alpha}{2}}}, \dfrac{\sum_{i=1}^{n}(x_i - \mu)^2}{C^{(n-1)}_{\frac{\alpha}{2}}}\right)$ 3. p-value $> \alpha$	1. $C^{(n-1)}_{\frac{\alpha}{2}} =$ chi2.ppf$(df = n-1, q = \alpha/2)$, 2. $C^{(n-1)}_{1-\frac{\alpha}{2}} =$ chi2.ppf$(df = n-1, q = 1 - \alpha/2)$, 3. p-value
Асимптотическая гипотеза о матожидании	1. $X \sim \mathcal{F}$ 2. $D(x) = \sigma^2$ - известно 3. $n \to \infty$ $(n \gg 0)$	$\mu = \mu_0$	$\mu \neq \mu_0$	$z_p = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0,1)$	Не отвергаем на уровне значимости α, если 1. $z_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{\sigma}{\sqrt{n}}\right)$ 3. p-value $> \alpha$	1. $z_{1-\frac{\alpha}{2}} =$ norm.ppf$(q = 1 - \alpha/2)$, 2. p-value $= 2 \cdot \left(1 - \text{norm.cdf}(\text{abs}(z_p))\right)$
Асимптотическая гипотеза о матожидании	1. $X \sim \mathcal{F}$ 2. $D(x) = \sigma^2$ - неизвестно 3. $n \to \infty$ $(n \gg 0)$	$\mu = \mu_0$	$\mu \neq \mu_0$	$z_p = \dfrac{\overline{X} - \mu_0}{S_0/\sqrt{n}} \xrightarrow[n\to\infty]{d} N(0,1)$	Не отвергаем на уровне значимости α, если 1. $z_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{X} - z_{1-\frac{\alpha}{2}}\dfrac{S_0}{\sqrt{n}}, \overline{X} + z_{1-\frac{\alpha}{2}}\dfrac{S_0}{\sqrt{n}}\right)$ 3. p-value $> \alpha$	1. $z_{1-\frac{\alpha}{2}} =$ norm.ppf$(q = 1 - \alpha/2)$, 2. p-value $= 2 \cdot \left(1 - \text{norm.cdf}(\text{abs}(z_p))\right)$
Bootstrap	1. $X \sim \mathcal{F}$ 2. n - небольшое	$\mu = \mu_0$ или $\sigma^2 = \sigma_0^2$	$\mu \neq \mu_0$ или $\sigma^2 \neq \sigma_0^2$	Генерируем много выборок из данной одинаковой длины. Считаем для каждой из них нужную статистику $\left(\overline{X}_i \text{ или } \widehat{\sigma}_i^2\right)$. Считаем квантили $q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}$ для выборки этих статистик.	Не отвергаем на уровне значимости α, если 1. μ_0 $(\sigma_0^2) \in \left(q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}}\right)$	1. scipy.stats.bootstrap 2. numpy.random.choice

2. Два распределения

Название	Предпосылки	H_0	H_1	Статистика	Выводы	Python (numpy, scipy.stats)
Гипотеза о разности матожиданий связанных пар	1. $X \sim N(\mu_x, \sigma_x^2)$ 2. $Y \sim N(\mu_y, \sigma_y^2)$ 3. $n = n_x = n_y$ 4. σ_x^2, σ_y^2 – известно	$\mu_x - \mu_y = \mu_0$	$\mu_x - \mu_y \neq \mu_0$	$\Delta = X - Y$, $z_p = \dfrac{\overline{\Delta} - \mu_0}{D(\overline{\Delta})} \sim N(0,1)$, $z_p = \dfrac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2 + \sigma_y^2}{n}}} \sim N(0,1)$	Не отвергаем на уровне значимости α, если 1. $z_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{\Delta} - z_{1-\frac{\alpha}{2}}D(\overline{\Delta}), \overline{\Delta} + z_{1-\frac{\alpha}{2}}D(\overline{\Delta})\right)$ 3. p-value $> \alpha$	См. аналогичное выше, p-value
Гипотеза о разности матожиданий связанных пар	1. $X \sim N(\mu_x, \sigma_x^2)$ 2. $Y \sim N(\mu_y, \sigma_y^2)$ 3. $n = n_x = n_y$ 4. σ_x^2, σ_y^2 – неизвестно	$\mu_x - \mu_y = \mu_0$	$\mu_x - \mu_y \neq \mu_0$	$\Delta = X - Y$, $t_p = \dfrac{\overline{\Delta} - \mu_0}{S_0(\Delta)/\sqrt{n}} \sim T_{n-1}$	Не отвергаем на уровне значимости α, если 1. $t_p \in \left(-t^{(n-1)}_{1-\frac{\alpha}{2}}, t^{(n-1)}_{1-\frac{\alpha}{2}}\right)$, 2. $\mu_0 \in \left(\overline{\Delta} \pm t^{(n-1)}_{1-\frac{\alpha}{2}}\dfrac{S_0(\Delta)}{\sqrt{n}}\right)$ 3. p-value $> \alpha$	См. аналогичное выше, p-value
Гипотеза о разности матожиданий	1. $X \sim N(\mu_x, \sigma_x^2)$ 2. $Y \sim N(\mu_y, \sigma_y^2)$ 3. $n_x \neq n_y$ 4. σ_x^2, σ_y^2 – известно	$\mu_x - \mu_y = \mu_0$	$\mu_x - \mu_y \neq \mu_0$	$z_p = \dfrac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}} \sim N(0,1)$	Не отвергаем на уровне значимости α, если 1. $z_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. $(\mu_x - \mu_y) \in \left(\overline{X} - \overline{Y} \pm z_{1-\frac{\alpha}{2}}\sqrt{\dfrac{\sigma_x^2}{n_x} + \dfrac{\sigma_y^2}{n_y}}\right)$ 3. p-value $> \alpha$	См. аналогичное выше, p-value

Название	Предпосылки	H_0	H_1	Статистика	Выводы	Python (numpy, scipy.stats)
Гипотеза о разности матожиданий	1. $X \sim N(\mu_x, \sigma^2)$ 2. $Y \sim N(\mu_y, \sigma^2)$ 3. $n_x \neq n_y$ 4. $\sigma^2 = \sigma_x^2 = \sigma_y^2$ – неизвестно	$\mu_x - \mu_y = \mu_0$	$\mu_x - \mu_y \neq \mu_0$	$\widehat{\sigma}^2 = \dfrac{S_0^2(X)(n_x-1) + S_0^2(Y)(n_y-1)}{n_x+n_y-2}$ $t_p = \dfrac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\widehat{\sigma}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} \sim T_{n_x+n_y-2}$	Не отвергаем на уровне значимости α, если 1. $t_p \in \left(-t_{1-\frac{\alpha}{2}}^{(n_x+n_y-2)}, t_{1-\frac{\alpha}{2}}^{(n_x+n_y-2)}\right)$, 2. $(\mu_x - \mu_y) \in \left(\overline{X} - \overline{Y} \pm t_{1-\frac{\alpha}{2}}^{(n_x+n_y-2)}\hat{\sigma}\right)$ 3. p-value $> \alpha$	1. $t_{1-\frac{\alpha}{2}}^{(n_x+n_y-2)} =$ t.ppf$(q = 1 - \alpha/2, df = n_x + n_y - 2)$, 2. p-value $= 2 \cdot \big(1 -$ t.cdf(abs$(t_p), df = n_x + n_y - 2)\big)$, 3. scipy.stats.ttest_ind
Гипотеза о *равенстве* матожиданий. Тест Уэлча	1. $X \sim N(\mu_x, \sigma_x^2)$ 2. $Y \sim N(\mu_y, \sigma_y^2)$ 3. $n_x \neq n_y$ 4. σ_x^2, σ_y^2 – неизвестно	$\mu_x - \mu_y = 0$	$\mu_x - \mu_y \neq 0$	$t_p = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{\widehat{\sigma_x^2}}{n_x} + \dfrac{\widehat{\sigma_y^2}}{n_y}}} \sim T_{\widehat{d}}$ $\widehat{d} = \dfrac{\left(\dfrac{\widehat{\sigma_x^2}}{n_x} + \dfrac{\widehat{\sigma_y^2}}{n_y}\right)^2}{\dfrac{\widehat{\sigma}_x^4}{n_x^2(n_x-1)} + \dfrac{\widehat{\sigma}_y^4}{n_y^2(n_y-1)}}$	Не отвергаем на уровне значимости α, если 1. $t_p \in \left(-t_{1-\frac{\alpha}{2}}^{(\widehat{d})}, t_{1-\frac{\alpha}{2}}^{(\widehat{d})}\right)$, 2. $0 \in \left(\overline{X} - \overline{Y} \pm t_{1-\frac{\alpha}{2}}^{(\widehat{d})}\sqrt{\dfrac{\widehat{\sigma_x^2}}{n_x} + \dfrac{\widehat{\sigma_y^2}}{n_y}}\right)$ 3. p-value $> \alpha$	1. $t_{1-\frac{\alpha}{2}} =$ t.ppf$(q = 1 - \alpha/2, df = d)$, 2. p-value $= 2 \cdot \big(1 -$ t.cdf(abs$(t_p), df = d)\big)$ 3. scipy.stats.ttest_ind(equal_var=Flase)
Гипотеза об отношении дисперсий	1. $X \sim N(\mu_x, \sigma_x^2)$ 2. $Y \sim N(\mu_y, \sigma_y^2)$ 3. $n_x \neq n_y$ 4. σ_x^2, σ_y^2 – неизвестно	$\dfrac{\sigma_x^2}{\sigma_y^2} = 1$	$\dfrac{\sigma_x^2}{\sigma_y^2} \neq 1$	$f_p = \dfrac{\widehat{\sigma_x^2}}{\widehat{\sigma_y^2}} \sim F_{n_x-1, n_y-1}$	Не отвергаем на уровне значимости α, если 1. $f_p \in \left(f_{\frac{\alpha}{2}}^{(n_x-1,n_y-1)}, f_{1-\frac{\alpha}{2}}^{(n_x-1,n_y-1)}\right)$, 2. p-value $> \alpha$	1. $f_{\frac{\alpha}{2}}^{(n_x-1,n_y-1)} =$ f.ppf$(dfn = n_x - 1, dfd = n_y - 1, q = a/2)$, 2. $f_{1-\frac{\alpha}{2}}^{(n_x-1,n_y-1)} =$ f.ppf$(dfn = n_x - 1, dfd = n_y - 1, q = 1 - a/2)$, 3. p-value

3. Критерии сравнения

Название	Предпосылки	H_0	H_1	Статистика	Выводы	Python (numpy, scipy.stats)
Критерий Пирсона (χ^2) о согласии	1. $X \sim \mathcal{F}_x$ 2. $\mathcal{F}_0$ – дискретное.	$\mathcal{F}_x = \mathcal{F}_0$	$\mathcal{F}_x \neq \mathcal{F}_0$	Для каждого значения a_i имеем частоту/количество (ν_i) в данной выборке и теоретическую вероятность p_i. $\rho = \sum\limits_{i=1}^{k} \dfrac{(\nu_i - np_i)^2}{np_i} \xrightarrow[n\to\infty]{H_0} \chi_{k-1}^2$	Не отвергаем на уровне значимости α, если 1. p-value $> \alpha$	1. p-value $= 2 \cdot$ chi2.cdf$(\rho, df = n - 1)$
Критерий Колмогорова о согласии	1. $X \sim \mathcal{F}_x$ 2. $\mathcal{F}_0$ – непрерывное.	$\mathcal{F}_x = \mathcal{F}_0$	$\mathcal{F}_x \neq \mathcal{F}_0$	$\widehat{F}_n(x)$ – эмпирическая функция распределения, $F_0(x)$ – функция распределения $\mathcal{F}_0$. $D_n = \sup\limits_x \left\lvert\widehat{F}_n(x) - F_0(x)\right\rvert$, $k_p = \sqrt{n}D_n \xrightarrow[n\to\infty]{d} \eta \sim \mathcal{K}(y)$ – функция распределения Колмогорова.	Не отвергаем на уровне значимости α, если 1. $k_p \leq K_{1-\alpha}$, 2. p-value $> \alpha$	1. scipy.stats.ksone 2. scipy.stats.ks_1samp 3. scipy.stats.kstest
Критерий Колмогорова-Смирнова об однородности	1. $X \sim \mathcal{F}_x$ 2. $Y \sim \mathcal{F}_y$	$\mathcal{F}_x = \mathcal{F}_y$	$\mathcal{F}_x \neq \mathcal{F}_y$	$\widehat{F}_{n_x}(x), \widehat{F}_{n_y}(x)$ – эмпирические функции распределения. $ks_p = \sqrt{\dfrac{n_x n_y}{n_x + n_y}} \sup\limits_x \left\lvert\widehat{F}_{n_x}(x) - \widehat{F}_{n_y}(x)\right\rvert$ $ks_p \xrightarrow[n_x,n_y\to\infty]{d} \eta \sim \mathcal{K}(y)$ – функция распределения Колмогорова.	Не отвергаем на уровне значимости α, если 1. $ks_p \leq K_{1-\alpha}$, 2. p-value $> \alpha$	1. scipy.stats.ksone 2. scipy.stats.ks_2samp
Критерий Пирсона (χ^2) о независимости	Объекты имеют пары из категорий (x_i, y_i). Всего X имеет s категорий, Y имеет k категорий.	X, Y - независимые	X, Y - зависимые	ν_{ij} - частоты пары категорий $(a_i, b_j) \sim (X, Y)$. $n_{i\bullet} = \sum\limits_{j=1}^{k} \nu_{ij},\ n_{\bullet j} = \sum\limits_{i=1}^{s} \nu_{ij}$ $\gamma = \sum\limits_{i=1}^{s}\sum\limits_{j=1}^{k} \dfrac{\left(\nu_{ij} - \dfrac{n_{i\bullet}n_{\bullet j}}{n}\right)^2}{\dfrac{n_{i\bullet}n_{\bullet j}}{n}} \sim \chi_{(s-1)(k-1)}^2$	Не отвергаем на уровне значимости α, если 1. $\gamma \in (0, C_{1-\alpha}^{(s-1)(k-1)})$, 2. p-value $> \alpha$	1. scipy.stats.contingency.crosstab 2. pandas.crosstab 3. scipy.stats.chi2_contingency (correction = False)
Коэффициент корреляции Спирмена	Объекты имеют пары из порядковых (ранговых) переменных (r_i, k_i).	X, Y - независимые	X, Y - зависимые	$S = \sum\limits_{i=1}^{n}(r_i - k_i)^2 \in \left[0, \dfrac{n^3 - n}{3}\right]$ $\rho = 1 - \dfrac{6S}{n^3 - n} \in [-1, 1]$ $\rho_p = \sqrt{n-1}\rho \xrightarrow[H_0]{n\to\infty} N(0,1)$	Не отвергаем на уровне значимости α, если 1. $\rho_p \in \left(-z_{1-\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}\right)$, 2. p-value $> \alpha$	1. scipy.stats.spearmanr 2. p-value

Красный текст – ссылка в этом документе. Синий текст – ссылка на страницу в интернете.