Pravdepodobnosť a štatistika pre informatikov

Riešené príklady
časť 2: štatistika

Beáta Stehlíková

Fakulta matematiky, fyziky a informatiky
Univerzita Komenského v Bratislave

december 2025

1 Chí-kvadrát test dobrej zhody

Príklad 1. Hádzali sme 600 krát klasickou hracou kockou. Jednotka padla 120 krát, dvojka 90 krát, trojka 89 krát, štvorka 93 krát, päťka 105 krát, šestka 103 krát. Chceme testovať hypotézu, že kocka je pravidelná chí-kvadrát testom dobrej zhody.

(a) Zostrojte tabuľku pozorovaných a očakávaných početností a pomocou nej vypočítajte testovaciu štatistiku.

(b) Pre aké hodnoty štatistiky sa hypotéza zamieta? Veľmi veľké, veľmi malé alebo pre obdive možnosti? Prečo?

(c) Aké je pravdepodobnostné rozdelenie štatistiky za platnosti nulovej hypotézy?

(d) Ako vypočítame kritickú hodnotu, určujúcu, či sa hypotéza zamietne alebo nie? Ako bude vyzerať rozhodovanie o zamietnutí?

(e) Ako vypočítame p-hodnotu? Ako bude vyzerať rozhodovanie o zamietnutí?

Riešenie. Testovanie pravdelnosti kocky je príklad, na ktorom bol vysvetľovaný na prednáške, takže postupujeme analogicky:

(a) Ak je kocka pravidelná, pravdepodobnosť každého počtu bodiek je $\frac{1}{6}$. Pri 600 pokusoch je očakávaná početnosť každej možnosti rovná $\frac{1}{6} \cdot 600 = 100$. Spolu s pozorovanými počtami zo zadania tak dostávame prvé tri stĺpce tabuľky 1, ktoré sú požadované v zadaní úlohy.

číslo na kocke	pozorovaná početnosť	očakávaná početnosť	rozdiel	sčítanec v sume
1	120	100	20	$\frac{20^2}{100}$
2	90	100	-10	$\frac{(-10)^2}{100}$
3	89	100	-11	$\frac{(-11)^2}{100}$
4	93	100	-7	$\frac{(-7)^2}{100}$
5	105	100	5	$\frac{5^2}{100}$
6	103	100	3	$\frac{3^2}{100}$

Tabuľka 1: Tabuľka k príkladu 1

Zostávajúce dva stĺpce nám pomôžu pri konštrukcii testovacej štatistiky. Poznamenajme ešte, že je jedno, či rozdiely počítame ako „očakávané mínus pozorované“ alebo „pozorované mínus očakávané“, lebo v nasledujúcom kroku sa umocnia na druhú. Hodnota štatistiky - označíme ju ako χ^2 (chí-kvadrát) je

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} = \frac{(-20)^2}{100} + \frac{(-10)^2}{100} + \frac{(-11)^2}{100} + \frac{(-7)^2}{100} + \frac{5^2}{100} + \frac{3^2}{100} = \frac{704}{100} = 7{,}04,$$

kde O_i sú pozorované počty (observed) a E_i sú očakávané počty (expected). Sčitujeme teda hodnoty v poslednom stĺpci tabuľky 1. Hodnota testovacej štatistiky teda je 7,04.

(b) Hodnota štatistiky je vždy nezáporná, nulová hodnota by znamenala dokonalú zhodu medzi očakávanými a pozorovanými počtami jednotlivých možností. Veľká hodnota znamená, že sa očakávané a pozorované hodnoty výrazne líšia. Nulovú hypotézu teda zamietneme pre veľmi veľké hodnoty testovacej štatistiky.

(c) Za platnosti nulovej hypotézy - ak neodhadujeme žiadne parametre rozdelenia, čo je aj náš prípad - platí, že štatistika má χ^2 rozdelenie s $k-1$ stupňami voľnosti, kde k je počet kategórií, do ktorých triedime dáta (teda počet riadkov v tabuľke). V našom prípade je $k = 6$, teda testovacia štatistika má za platnosti nulovej hypotézy χ^2 rozdelenie s 5 stupňami voľnosti.

(d) Test má fungovať tak, že ak nulová hypotéza platí, tak ju zamietneme so stanovenou pravdepodobnosťou α, ktorá sa nazýva hladina významnosti. Uvažujme štandardnú hladinu významnosti $\alpha = 0{,}05$. Keďže hypotézu zamietame pre veľké hodnoty štatistiky, kritickú hodnotu nájdeme ako takú hodnotu $\chi^2_{\text{kritická}}$, pre ktorú je $\mathbb{P}(\chi^2 \geq \chi^2_{\text{kritická}}) = 0{,}05$, ak X je náhodná premenná s χ^2 rozdelenie s 5 stupňami voľnosti.

Potom pre vypočítanú hodnotu štatitiky χ^2 funguje rozhodovanie nasledovne: Ak $\chi^2 \geq \chi^2_{\text{kritická}}$, nulovú hypotézu zamietame. Ak $\chi^2 < \chi^2_{\text{kritická}}$, nulovú hypotézu nezamietame. (Je jedno, kam dáme znamienko rovnosti.)

V našom prípade $\chi^2 = 7{,}04$ a kritická hodnota je podľa výpočtu na obrázku 1 rovná $\chi^2 = 11{,}07$, teda nezamietame nulovú hypotézu o pravidelnej kocke.

```python
from scipy.stats import chi2

alpha = 0.05
df = 5
chi2_krit = chi2.ppf(1 - alpha, df)
print(f"kritická hodnota: {critical_value:.2f}")
```

```
kritická hodnota: 11.07
```

Obr. 1: Výpočet kritickej hodnoty v príklade 1. Hľadáme hodnotu $\chi^2_{\text{kritická}}$, pre ktorú je $\mathbb{P}(X \geq \chi^2_{\text{kritická}}) = 0{,}05$,. Teda ak F je distribučná funkcia, tak $F(\chi^2_{\text{kritická}}) = \mathbb{P}(X \leq \chi^2_{\text{kritická}}) = 0{,}95$. Potrebujeme teda hodnotu inverznej funkcie k distribučnej funkcii v bode 0,95.

(e) P-hodnota je definovaná ako pravdepodobnosť, že za platnosti nulovej hypotézy bude testovacia štatistika χ^2 rovnaká alebo „horšia “väčšia ako vypočítaná hodnota. Teda ak X je je náhodná premenná s χ^2 rozdelenie s 5 stupňami voľnosti, potrebujeme $\mathbb{P}(X \geq \chi^2)$.

P-hodnotu porovnávame s hladinou významnosti. Uvažujme štandardnú hladinu významnosti $\alpha = 0{,}05$. Ak je p-hodnota menšia ako 0,05, tak nulovú

hypotézu zamietame. Ak je p-hodnota väčšia alebo rovná 0,05, tak nulovú hypotézu nezamietame.

V našom prípade je p-hodnota na základe výpočtu na obrázku 2 rovná 0,2177, teda nezamietame nulovú hypotézu o pravidelnej kocke.

```
from scipy.stats import chi2

statistika = 7.04
df = 5
p_value = 1 - chi2.cdf(statistika, df)
print(f"p-hodnota: {p_value:.4f}")
```

```
p-hodnota: 0.2177
```

Obr. 2: Výpočet p-hodnoty v príklade 1. Pre vypočítanú hodnotu štatistiky χ^2 Hľadáme pravdepodobnosť $\mathbb{P}(X \geq \chi^2)$. Teda ak F je distribučná funkcia, tak potrebujeme $1 - F(\chi^2)$

Poznámka. Na obrázku 3 uvádzame testovanie tejto hypotézy pomocou funkcie zabudovanej v Pythone. Vidíme, že hodnota štatistiky aj p-hodnota sa zhoduje s našim výpočtom.

```
from scipy.stats import chisquare

pozorovane = [120, 90, 89, 93, 105, 103]
ocakavane = [100]*6

stat, pval = chisquare(f_obs=pozorovane, f_exp=ocakavane)
print(f"štatistika: {stat:.4f}")
print(f"p-hodnota: {pval:.4f}")
```

```
štatistika: 7.0400
p-hodnota: 0.2177
```

Obr. 3: Testovanie hypotézy z príkladu 1 pomocou funkcie zabudovanej v Pythone.

Príklad 2. Máme dve mince a vznikla otázka, či sú obe pravidelné. Preto nimi hádzali spolu 200 krát, pričom v 60 prípadoch padli dve hlavy, v 44 prípadoch padli dva znaky a v 96 prípadoch padol jeden znak a jedna hlava. Na základe týchto údajov, máme rozhodnúť, či sa danú dve mince môžu považovať za pravidelné alebo nie. Chceme pritom použiť chí-kvadrát test dobrej zhody.

(a) Vypočítajte testovaciu štatistiku.

(b) Aké je pravdepodobnostné rozdelenie štatistiky za platnosti nulovej hypotézy?

(c) Ktorá z hodnôt na obrázku (c) je správnou kritickou hodnotou pre tento test, ak chceme použiť hladinu významnosti 0,05? Ako bude vyzerať rozhodovanie o zamietnutí?

(d) Zamietama v tomto prípade hypotézu o pravidelných minciach alebo nie?

(e) Čo vieme bez ďalšieho výpočtu povedať o p-hodnote? Je väčšia alebo menšia ako 0,05

```
from scipy.stats import chi2

A = chi2.ppf(0.05, df=4)
B = chi2.ppf(0.95, df=4)
C = chi2.ppf(0.05, df=3)
D = chi2.ppf(0.95, df=3)
E = chi2.ppf(0.05, df=2)
F = chi2.ppf(0.95, df=2)

for i in [A, B, C, D, E, F]:
  print(round(i, 5))
```

```
0.71072
9.48773
0.35185
7.81473
0.10259
5.99146
```

Obr. 4: Hodnoty k príkladu 2.

Riešenie.

(a) Máme 200 hodov, pozorované počty sú: $O_{\text{HH}} = 60$, $O_{\text{ZZ}} = 44$, $O_{\text{HZ, ZH}} = 96$. Predpoklad nulovej hypotézy je, že obe mince sú pravidelné. Hody nezávislé, takže pravdepodobnosti jednotlivých výsledkov sú:

$$P(\text{HH}) = 0{,}5 \cdot 0{,}5 = 0{,}25, \quad P(\text{ZZ}) = 0{,}5 \cdot 0{,}5 = 0{,}25,$$

$$P(\text{HZ alebo ZH}) = 0{,}5 \cdot 0{,}5 + 0{,}5 \cdot 0{,}5 = 0{,}5$$

Očakávané početnosti týchto výsledkov pri 200 pokusoch potom sú:

$$E_{\text{HH}} = 0.25 \cdot 200 = 50, \quad E_{\text{ZZ}} = 0.25 \cdot 200 = 50, \quad E_{\text{HZ, ZH}} = 0.5 \cdot 200 = 100$$

Testovacia štatistika je

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(60-50)^2}{50} + \frac{(44-50)^2}{50} + \frac{(96-100)^2}{100} = \frac{100}{50} + \frac{36}{50} + \frac{16}{100} = 2{,}88$$

(b) Kategórie, pre ktoré sa porovnávajú pozorované a očakávané počty, sú tri. Takže pravdepodobnostné rozdelenie štatistiky za platnosti nulovej hypotézy je chí-kvadrát s dvoma stupňami voľnosti.

(c) Hypotéza sa zamietne, ak je štatistika väčšia ako kritická hodnota. Štatistika má za platnosti nulovej hypotézy je chí-kvadrát s dvoma stupňami voľnosti. Kritická hodnota je teda také číslo x, pre ktoré je pravdepodobnosť, chí-kvadrát s dvoma stupňami voľnosti nadobudne hodnotu väčšiu ako x, rovná 0,05

Ak F je distribučná funkcia chí-kvadrát s dvoma stupňami voľnosti, tak $F(x) = \mathbb{P}(X \leq \chi^2_{\text{kritická}}) = 0{,}95$. Potrebujeme teda hodnotu inverznej funkcie k distribučnej funkcii v bode 0,95. V Pythone je táto inverzná funkcia pod názvom `ppf` a parameter `df`, označujúci počet stupňov voľnosti, je rovný 2. Z čísel na obrázku je kritickou hodnou číslo F.

(d) Hodnota testovacje štatistiky je 2,88. Nulová hypotéza sa zamieta pre hodnoty štatistiky, ktoré sú väčšie ako kritická hdonota, pričom naša kritická hodnota je 5,99146. Nulovú hypotézu o pravidelných minciach teda nezamietame.

(e) Na hladine významnosti 0,05 sme hypotézu nezamietli. To znamená, že p-hodnota je väčšia ako 0,05.

Príklad 3. Hádžeme šiestimi kockami a zaznamenávame, koľkokrát padla šestka. Žiadna šestka nepadla v 345 pokusov, jedna šestka padla v 411 pokusoch, dve v 187 pokusoch, tri v 50 pokusoch, štyri v 6 pokusoch, päť v zostávajúcom jednom pokuse. Ani raz sa nestalo, že by na všetkých kockách padli šestky. Na základe týchto údajov, máme rozhodnúť, či sú dané kocky pravidelné alebo nie. Chceme pritom použiť chí-kvadrát test dobrej zhody. Začiatok výpočtu je na obrázku 5.

```python
from scipy.stats import binom

n_hody = 1000    # pocet pokusov
n_kocky = 6      # pocet kociek v jednom pokuse
p_sestka = 1/6   # pravdepodobnost sestky

print(f"{'počet šestiek':>15} | {'pravdepodobnosť':>15} | {'očakávaná početnosť':>20}")
print("-"*60)

for i in range(n_kocky + 1):
    p = binom.pmf(i, n_kocky, p_sestka)
    expected = n_hody * p
    print(f"{i:>15} | {p:>15.5f} | {expected:>20.2f}")
```

```
  počet šestiek | pravdepodobnosť |  očakávaná početnosť
------------------------------------------------------------
              0 |         0.33490 |               334.90
              1 |         0.40188 |               401.88
              2 |         0.20094 |               200.94
              3 |         0.05358 |                53.58
              4 |         0.00804 |                 8.04
              5 |         0.00064 |                 0.64
              6 |         0.00002 |                 0.02
```

Obr. 5: Výpočet k príkladu 3.

(a) Vysvetlite, ako boli počítané pravdepodobnosti `p` vo for-cykle.

(b) Vysvetlite, akú úpravu ešte musíme spraviť pre tým, ako začneme počítať testovaciu štatistiku porovnávaním očakávaných a pozorovaných početností.

(c) Pomocou zaokrúhlených hodnôt z tabuľky vypočítajte hodnotu testovacej štatistiky (stačí súčet bez vyčíslenia).

(d) Aké je pravdepodobnostné rozdelenie štatistiky za platnosti nulovej hypotézy?

Riešenie.

(a) Počíta sa pravdepodobnosť toho, že pri hádzaní šiestimi kockami padne F šestiek. Vypočítame ju pomocou binomického rozdelnia, ide o $\mathbb{P}(X = i)$, kde X má binomické rozdelenie s parametrami $n = 6$ a $p = \frac{1}{6}$.

(b) Podmienkou použitia testu je, že očakávané početnosti sú aspoň 5. Preto možnosti 4, 5, 6 šestiek zlúčime do jednej - „4 a viac šestiek“. Pri použití zaokrúhlených čísel v tabuľke[1] bude očakávaná početnosť rovná $8{,}04 + 0{,}64 + 0{,}02 = 8{,}70$.

(c) Testovacia štatistika je

$$\chi^2 = \frac{(345 - 334{,}90)^2}{334{,}90} + \frac{(411 - 401{,}88)^2}{401{,}88} + \frac{(187 - 200{,}94)^2}{200{,}94} + \frac{(50 - 53{,}58)^2}{53{,}58} + \frac{(7 - 8{,}70)^2}{8{,}70}.$$

(d) Po zlúčení možností s malými pravdepodobnosťmi v bode (b) uvažujeme päť výsledkov, pre ktoré porovnávame očakávané a pozorované početnosti. Takže pravdepodobnostné rozdelenie je chí-kvadrát rozdelenie so štyrmi stupňami voľnosti.

Príklad 4. V banke zaznamenávali počet zákazníkov, ktorí prišli na pobočku počas desaťminútových intervalov. Počty intervalov s jednotlivými počtami zákazníkov sú dané v tabuľke 2. Chceme zistiť, či môžeme pravdepodobnostné rozdelenie týchto dát považovať za Poissonovo. Použijeme pritom chí-kvadrát test dobrej zhody.

počet zákazníkov	0	1	2	3	4
počet intervalov	24	33	17	3	3

Tabuľka 2: Tabuľka k príkladu 4

(a) Odhadnite parameter Poissonovho rozdelenia.

(b) Zlúčte možnosti „traja zákazníci“ a „štyria zákazníci“ do jednej - „traja a viacerí“. Ktorý z výpočtov na obrázku 6 dáva správny výpočet očakávaných početností, ak v premennej `lam` je odhad parametra λ a v premmenej N je počet dát? Zdôvodnite svoju odpoveď.

(c) Aké je pravdepodobnostné rozdelenie štatistiky za platnosti nulovej hypotézy?

```
expected_A = np.array([
    poisson.pmf(0, lam),
    poisson.pmf(1, lam),
    poisson.pmf(2, lam),
    1 - poisson.cdf(2, lam)]) * N
```

```
expected_B = np.array([
    poisson.pmf(0, lam),
    poisson.pmf(1, lam),
    poisson.pmf(2, lam),
    poisson.pmf(3, lam)])
```

```
expected_C = np.array([
    poisson.pmf(0, lam),
    poisson.pmf(1, lam),
    poisson.pmf(2, lam),
    1 - poisson.cdf(3, lam)]) * N
```

```
expected_D = np.array([
    poisson.cdf(0, lam),
    poisson.cdf(1, lam),
    poisson.cdf(2, lam),
    poisson.cdf(3, lam)]) * N
```

Obr. 6: Výpočet k príkladu 4.

```
from scipy.stats import chisquare

data  = np.array([24, 33, 17, 6])
chisquare(f_obs=data, f_exp=expected)
```

```
Power_divergenceResult(statistic=np.float64(1.8177143241875464), pvalue=np.float64(0.6110877083111048))
```

```
data  = np.array([24, 33, 17, 6])
chisquare(f_obs=data, f_exp=expected, ddof=1)
```

```
Power_divergenceResult(statistic=np.float64(1.8177143241875464), pvalue=np.float64(0.402984506939315))
```

```
data  = np.array([24, 33, 17, 6])
chisquare(f_obs=data, f_exp=expected, ddof=2)
```

```
Power_divergenceResult(statistic=np.float64(1.8177143241875464), pvalue=np.float64(0.17758558804800917))
```

Obr. 7: Výpočet k príkladu 4.

(d) Ktorý z výpočtov na obrázku 7 je správnym použitím chí-kvadrát testu dobrej zhody a prečo? Aký je výsledok testu - zamietame hypotézu o Poissonovom rozdelení alebo nie?

Riešenie.

(a) Strednou hodnotou Poissonovho rozdelenia je jeho parameter λ, odhadneme ho preto ako aritmetický priemer dát. Celkovo je $24 + 33 + 17 + 3 + 3 = 80$, a teda ich priemer je

$$\lambda = \frac{1}{80}\left(0 \cdot +1 \cdot 33 + 2 \cdot 17 + 3 \cdot 3 + 4 \cdot 3\right) = \frac{88}{80} = \frac{11}{10} = 1{,}1.$$

(b) Ak X je náhodná premenná s naším Poissonovým rozdelením, tak požadované početnosti sú $N \cdot \mathbb{P}(X = 0), N \cdot \mathbb{P}(X = 1), N \cdot \mathbb{P}(X = 2), N \cdot \mathbb{P}(X \geq 3)$. Vieme (alebo si nájdeme vo vzorcovníku), že v Pythone `pmf` počíta *probability mass function* - teda pravdepodobnosti daných hodnôt a `cdf` počíta *cummulative distribution function* distribučnú funkciu. Takže ak sa pozrieme na obrázok 6:

[1] prakticky by sme vypočítali presnú hodnotu

- Pravdepodobnosti musia byť vynásobené počtom pozorovaní N, aby sa z nich stali očakávané početnosti. Teda možnosť B nevyhovuje, lebo sú tam len pravdepodobnosti (navyše posledná nie je správna).
- Prvé tri hodnoty musia byť `N*pmf(0, lam), N*pmf(1, lam), N*pmf(2, lam)`, teda možnosť D nevyhovuje, lebo tam sú namiesto pravdepodobností hodnoty distribučnej funkcie.
- Zostali možnosti A a C, ktoré sa líšia poslednou hodnotou. Príslušné pravdepodobnosti, ak označíme F distribučnú funkciu, sú

$$p_A = 1 - F(2) = 1 - \mathbb{P}(X \leq 2) = \mathbb{P}(X > 2) = \mathbb{P}(X \geq 3),$$

$$p_C = 1 - F(3) = 1 - \mathbb{P}(X \leq 3) = \mathbb{P}(X > 3) = \mathbb{P}(X \geq 4).$$

Správna je teda možnosť A[2].

(c) Štatistika má chí-kvadrát rozdelenie. Máme štyri kategórie hodnôt, do ktorých sú zatriedené dáta, takže ak by sme neodhadovali parametre rozdelenia, štatistika by mala tri stupne voľnosti. Keďže však odhadujeme jeden parameter, počet stupňov voľnosti sa zníži o jeden. Takže štatistika má chí-kvadrát rozdelenie s dvoma stupňami voľnosti.

(d) Parameter `ddof` vyjadruje *delta degrees of freedom*, teda o koľko sa zníži počet stupňov voľnosti kvôli odhadovaniu parametrov. V našom prípade o 1 (ako sme to napísali v predchádzajúcom bode), teda správny výstup je ten, ktorý pochádza z volania funkciu s parametrom `ddof=1`. Príslušná p-hodnota je 0,403. To je viac ako 0,05, takže na štandardnej päťpercentnej hladine významnosti hypotézu o Poissonovom rozdelení nezamietame.

[2]Na písomke samozrejme stačí ukázať, že vami zvolená možnosť obsahuje správne hodnoty (nie je nutné vysvetliť, v čom sú ostatné možnosti nesprávne). Za uhádnutú odpoveď, resp. nesprávne zdôvodnenie však body nie sú, takže ak neviete určiť správnu odpoveď, je lepšie napísať (so správnym zdôvodnením), čo nie je správna možnosť. Za to sa čiastočný počet bodov získať dá.

2 Pravdepodobnosť úspechu v postupnosti nezávislých pokusov

Príklad 1. Vyučujúci zmenil systém skúšania, namiesto riešenia úloh bude skúška pozostávať z testu. V minulosti bola pravdepodobnosť úspešného absolvovania skúšky na prvom termíne 70 percent. Teraz z 80 študentov spravilo v prvom termíne skúšku 60. Je zrejmé, že kvôli náhodnosti nemôžeme ani pri úspešnosti 70 percent očakávať, že skúšku spraví *presne* 70 percent študentov. Môžeme odchýlku od 70 percent pripísať náhodnosti? Alebo je rozdiel významný a môžeme teda povedať, že zmenou typu skúšky sa zmenila úspešnosť?

(a) Sformulujte nulovú hypotézu a alternatívnu hypotézu.

(b) Predpokladajme, že platí nulová hypotéza. Na prvý termín skúšky prišlo 80 študentov. Aké je pravdepodobnostné rozdelenie počtu študentov, ktorí skúšku úspešne spravia?

(c) Odvoďte, prečo môžeme rozdelenie z predchádzajúceho bodu aproximovať normálnym. Vypočítajte parametre tohto rozdelenia.

(d) Ako nám pri rozdhodnutí o testovaní hypotézy pomôžu kvantily normalizovaného normálneho rozdelenia na obrázku 8? Odvoďte postup, pomocou ktorého sa rozhodne o zamietnutí alebo nezamietnutí nulovej hypotézy na hladine významnosti 5 percent.

```
q_0025 = norm.ppf(0.025)
q_0975 = norm.ppf(0.975)
print(f"2.5% kvantil: {q_0025:.5f}")
print(f"97.5% kvantil: {q_0975:.5f}")
```

```
2.5% kvantil: -1.95996
97.5% kvantil: 1.95996
```

Obr. 8: Výpočet k príkladu 1.

(e) Ako sa vypočíta p-hodnota testu? Vyjadrite ju pomocou distribučnej funkcie normalizovaného normálneho rozdelenia..

(f) Na základe výstupu z Pythonu na obrázku spravte záver - zamietame nulovú hypotézu alebo nie?

Riešenie. Ide o analógiu testovania pravdepodobnosti padnutia hlavy na minci z prednášky (príklad s euromincou, ktorý sa dostal do správ).

(a) Označme p pravdepodobnosť úspechu na prvom termíne po zmene skúšky. Chceme zistiť, či sa úspešnosť zmenila - mohla sa zlepšiť alebo zhoršiť. preto ide o obojstranný test:

$$H_0 : p = 0{,}7, \qquad H_1 : p \neq 0{,}7$$

```
from statsmodels.stats.proportion import proportions_ztest

p0 = 0.7
proportions_ztest(60, 80, value=p0, prop_var=p0)
```

```
(np.float64(0.975900072948534), np.float64(0.32911398597860764))
```

Obr. 9: Výpočet k príkladu 1.

(b) Výsledky skúšky jednotlivých študentov môžeme považovať za nezávislé. Preto ide o 80 nezávislých pokusov, pričom pravdepodobnosť úspechu v pokuse je 0,7. Počet študentov (označme ho X), ktoré spravili skúšku, je počet úspechov v tejto sérii pokusov, takže pravdepodobnostné rozdelenie tejto náhodnej premennej je $\text{Bin}(n, p)$, kde $n = 80$, $p = 0{,}7$.

(c) Pre veľký počet pokusov môžeme $\text{Bin}(n, p)$ na základe centrálnej limitnej vety aproximovať normálnym rozdelením s rovnakou strednou hodnotou a disperziou[3].

Podľa centrálnej limitnej vety má súčet nezávislých náhodných premenných s rovnakým rozdelením limitne normálne rozdelenie. Náhdonú premennú X s binomickým rozdelením $\text{Bin}(n, p)$ môžeme napísať ako súčet nezávislých náhodných premenných s rovnakým rozdelením $X_1 + \cdots + X_n$, kde X_i sa rovná 1 s pravdepodobnosťou p a 0 s pravdepodobnosťou $1 - p$ (teda X_i vyjadruje, či v i-tom pokuse nastal úspech).

Stredná hodnota a disperzia tohto binomického rozdelenia je

$$\mathbb{E}(X) = n \cdot p = 80 \cdot 0{,}7 = 56, \quad \mathbb{D}(X) = n \cdot p \cdot (1 - p) = 80 \cdot 0{,}7 \cdot 0{,}3 = 16{,}8.$$

teda ho aproximujeme normálnym rozdelením

$$X \overset{aprox}{\sim} \mathcal{N}(\mu, \sigma^2), \text{ kde } \mu = 56, \quad \sigma^2 = 16.8.$$

(d) Keďže alternatívna hypotéza hovorí, že pravdepodobnosť nie je 0,7, nulovú hypotézu zamietneme pre veľmi veľké aj veľmi malé hodnoty X. Hranice musia byť také, aby pravdepodobnosť zamietnutia pravdivej hypotézy bola 5 percent.

Na obrázku 8 vidíme, že pre normalizované normálne rozdelenie tieto hranice máme vypočítané: pre $Z \sim \mathcal{N}(0, 1)$ platí

$$\mathbb{P}(Z < -1{,}95996) = 0{,}025,$$

$$\mathbb{P}(Z < 1{,}95996) = 0{,}975 \Rightarrow \mathbb{P}(Z > 1{,}95996) = 0{,}025.$$

Takže ak by štatistika mala $Y \sim \mathcal{N}(0, 1)$ a chceli by sme obor zamietnutia v tvare „veľmi malé a veľmi veľké hodnoty“, znamenalo by to „menšie ako -1,95996 alebo väčšie ako 1,95996“. Takúto štatistiku však vieme z našej náhodnej premennej X vyrobiť:

$$X \sim \mathcal{N}(56; 16{,}8) \Rightarrow Z = \frac{X - 56}{\sqrt{16{,}8}} \sim \mathcal{N}(0, 1).$$

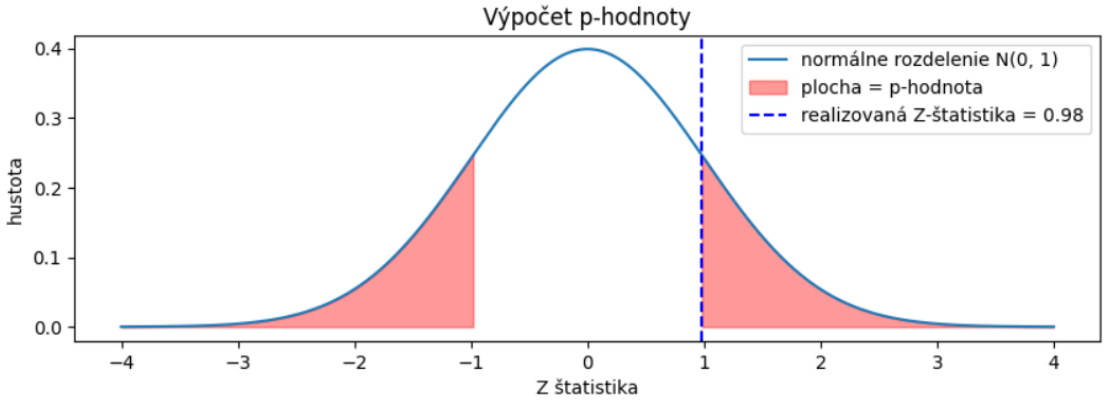
[3] špeciálny prípad prvého príkladu v prednáške o centrálnej limitnej vete

Vypočítame teda testovaciu štatistiku $Z = \frac{X-56}{\sqrt{16{,}8}}$, kde X je počet úspešných študentov z celkového počtu 80. Nulovú hypotézu zamietneme, ak je táto štatistika menšia ako -1,95996 alebo väčšia ako 1,95996.

(e) P-hodnota je pravdepodobnosť, že za platnosti nulovej hypotézy dostaneme štatistiku aspoň tak extrémnu ako je naša pozorovaná. V našom prípade je $X = 60$ a štatistika je rovná $Z = \frac{60-56}{\sqrt{16{,}8}} = \frac{4}{\sqrt{16{,}8}}$. Keďže máme obostrannú alternatívnu hypotézu, „aspoň tak extrémna hodnota“ znamená štatistika väčšia alebo rovná $\frac{4}{\sqrt{16{,}8}}$ alebo menšia alebo rovná $-\frac{4}{\sqrt{16{,}8}}$. To má pravdepodobnosť

$$\begin{aligned} p &= \mathbb{P}\left(Z \geq \frac{4}{\sqrt{16{,}8}}\right) + \mathbb{P}\left(Z \leq -\frac{4}{\sqrt{16{,}8}}\right) \\ &= 1 - \mathbb{P}\left(Z < \frac{4}{\sqrt{16{,}8}}\right) + \mathbb{P}\left(Z \leq -\frac{4}{\sqrt{16{,}8}}\right) \\ &= 1 - \Phi\left(\frac{4}{\sqrt{16{,}8}}\right) + \Phi\left(-\frac{4}{\sqrt{16{,}8}}\right), \end{aligned}$$

kde sme využili, že Z má za platnosti nulovej hypotézy normalizované normálne rozdelenie. Graficky je tento výpočet naznačený na obrázku 10.



Obr. 10: Výpočet p-hodnoty v príklade 1.

(f) P-hodnota na obrázku 9 je 0,32911, čo je viac ako 0,05, takže na 5-percentnej hladine významnosti nulovú hypotézu o úspešnosti 70 percent nezamietame.

Príklad 2. Kuriérska spoločnosť zoptimalizovala svoje trasy. Chce zistiť, či sa jej podarilo zlepšiť doručovanie. Doteraz bola pravdepodobnosť úspešného doručovania (pod čím rozumieme splnenie interných kritérií na dodržanie časov oznámených zákazníkom) v daný deň 0,8. Po zavedení nových trás bolo z 50 nezávislých dní doručovania úspešných 46.

(a) Sformulujte nulovú hypotézu a alternatívnu hypotézu.

(b) Predpokladajme, že platí nulová hypotéza. Aké je pravdepodobnostné rozdelenie počtu úspešných dní? Akým rozdelením ho môžeme aproximovať?

(c) Vypočítajte hodnotu Z-štatistiky pre tento test. Ako súvisí s rozdelením z predchádzajúceho bodu?

(d) Ako sa rozhodneme o zamietnutí alebo nezamietnutí nulovej hypotézy na hladine významnosti 5 percent?

(e) Ako sa vypočíta p-hodnota testu?

(f) Ktorý z výstupov z Pythonu na obrázku je relevantný pre náš príklad? Spravte na základe neho záver o testovanej hypotéze.

Riešenie.

(a) Označme p pravdepodobnosť úspešného doručenia kuriérom po zavedení nových trás. Chceme overiť, či sa úspešnosť zlepšila - preto ide o jednostranný test s alternatívou:

$$H_0 : p = 0{,}8, \qquad H_1 : p > 0{,}8$$

Dôvod takejto alternatívy je ten, že v prípade zamietnutia nulovej hypotézy chceme tvrdiť, že došlo k zlepšeniu.

(b) Každý deň doručovania môžeme považovať za nezávislý pokus. Počet úspešných dní X z $n = 50$ dní má binomické rozdelenie:

$$X \sim \mathrm{Bin}(n = 50, p = 0{,}8)$$

(c) Pre veľký počet pokusov môžeme X aproximovať normálnym rozdelením podľa centrálnej limitnej vety, pričom stredná hodnota a disperzia tohto normálneho rozdelenia sa rovná strednej hodnote a disperzii X:

$$X \overset{\text{aprox}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \mu = n \cdot p = 40, \quad \sigma^2 = n \cdot p \cdot (1-p) = 50 \cdot 0.8 \cdot 0.2 = 8,$$

(d) Testovacia štatistika Z je rovnako ako v predchádzajúcom príklade normalizovaná náhodná premenná X:

$$Z = \frac{X - \mu}{\sigma} = \frac{46 - 40}{\sqrt{8}} = \frac{6}{\sqrt{8}}$$

(e) Za platnosti nulovej hypotézy má Z normalizované normálne rozdelenie. Nulovú hypotézu zamietneme v prospech alternatívnej v prípade veľmi veľkého počtu úspešných dní X. To je ekvivalentné s tým, že štatistika Z nadobúda veľmi veľkú hodnotu. Hranica sa určí tak, aby pravdepodobnosť zamietnutia pravdivej nulovej hypotézy bola 0,05. Ak je nulová hypotéza pravdivá, tak Z má normalizované normálne rozdelenie, teda hľadaná hranica z je taká, pre ktorú

$$\begin{aligned} \mathbb{P}(\mathcal{N}(0,1) > z) &= 0{,}05, \\ \mathbb{P}(\mathcal{N}(0,1) \leq z) &= 0{,}95, \\ \Phi(z) &= 0{,}95, \\ z &= \Phi^{-1}(0{,}95). \end{aligned}$$

Hypotéza sa teda zamieta, ak je Z štatistika väčšia ako $\Phi^{-1}(0{,}95)$.

(f) P-hodnota pre jednostranný test sa vypočíta ako pravdepodobnosť, že za platnosti nulovej hypotézy bude štatistika aspoň taká extrémna ako je jej pozorovaná hodnota. Extrémna tu znamená, že sa odchyľuje od nulovej hypotézy v prospech alternatívnej. Teda ide o pravdepodobnosť, že počet úspešných dní bude taký, ako sme ho pozorovali alebo ešte väčší. V reči Z-štatistiky to znamená, že ide o pravdepodobnosť, že táto štatistika bude taká, ako sme ju pozorovali alebo väčšia:

$$p = \mathbb{P}\left(Z \geq \frac{6}{\sqrt{8}}\right) = 1 - \mathbb{P}\left(Z < \frac{6}{\sqrt{8}}\right) = 1 - \Phi\left(\frac{6}{\sqrt{8}}\right).$$

Príklad 4. Uvažujme výsledky písomky z príkladu 1.

(a) Zostrojte 90-percentný a 95-percentný interval spoľahlivosti ore pravdepodobnosť úspešného absolvovania skúšky. Použite pritom správne hodnoty z obrázka 11

(b) Predpokladajme, že chceme spraviť aj 99-percentný interval spoľahlivosti. Čo vieme povedať o jeho dĺžke v porovnaní s dĺžkou intervalov z prechádzajúceho bodu bez toho, aby sme ho konkrétne počítali?

```
from scipy.stats import norm

k = [0.005, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.995]
for i in k:
  print(f"P(N(0,1) <= {norm.ppf(i):.2f}) = {i}")
```

```
P(N(0,1) <= -2.58) = 0.005
P(N(0,1) <= -1.96) = 0.025
P(N(0,1) <= -1.64) = 0.05
P(N(0,1) <= -1.28) = 0.1
P(N(0,1) <= 1.28) = 0.9
P(N(0,1) <= 1.64) = 0.95
P(N(0,1) <= 1.96) = 0.975
P(N(0,1) <= 2.58) = 0.995
```

Obr. 11: Kvantily.

Riešenie.

(a) Vzorec pre interval spoľahlhivosti máme vo vzorcovníku:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

kde $n = 80$, $\hat{p} = \frac{60}{80} = 0{,}75$. Treba ešte určiť kvantily. Vzťah pre interval spoľahlivosti je pre $100(1-\alpha)$-percentný interval spoľahlivosti a hodnota z_q je také číslo, ktoré spĺňa podmienku $\mathbb{P}(\mathcal{N}(0,1) \leq z_q) = q$. Teda:

- V prípade 90-percentného intervalu spoľahlivosti je $\alpha = 0{,}1$, a teda potrebujeme hodnotu $z_{0{,}95}$, pre ktorú je $\mathbb{P}(\mathcal{N}(0,1) \leq z_{0{,}95}) = 0{,}95$. Z obrázku 11 vidíme, že $z_{0{,}95} = 1{,}64$. 90-percentný interval spoľahlivosti teda je

$$\left[0{,}75 - 1{,}64 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{80}}, 0{,}75 + 1{,}64 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{80}}\right].$$

- V prípade 95-percentného intervalu spoľahlivosti je $\alpha = 0{,}05$, a teda potrebujeme hodnotu $z_{0{,}975}$, pre ktorú je $\mathbb{P}(\mathcal{N}(0,1) \leq z_{0{,}975}) = 0{,}975$. Z obrázku 11 vidíme, že $z_{0{,}975} = 1{,}96$. 95-percentný interval spoľahlivosti teda je

$$\left[0{,}75 - 1{,}96 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{80}}, 0{,}75 + 1{,}96 \cdot \sqrt{\frac{0{,}75 \cdot 0{,}25}{80}}\right].$$

(b) 99-percentný interval spoľahlivosti obsahuje neznámy parameter p s pravdepodobnosťou 99 percent, 95 percentný s pravdeodobnosťou 95 percent, 90 percentný s pravdepodobnosťou 90 percent. S čím väčšou pravdepodobnosťou má interval obsahovať parameter p, tým je širší. Teda 99-percentný interval spoľahlivosti bude širší ako obidva intervaly z predchádzajúceho bodu.

Príklad 5. V prieskume verejnej mienky vyjadarilo určitému politikovi podporu 245 ľudí z 1000 oslovených. Zostrojte 95-percentný interval spoľahlivosti pre jeho podporu.

Riešenie. Vzorec pre interval spoľahlhivosti máme vo vzorcovníku:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

kde $n = 1000$, $\hat{p} = \frac{245}{1000} = 0{,}245$. Treba ešte určiť kvantily. Vzťah pre interval spoľahlivosti je pre $100(1-\alpha)$-percentný interval spoľahlivosti a hodnota z_q je také číslo, ktoré spĺňa podmienku $\mathbb{P}(\mathcal{N}(0,1) \leq z_q) = q$. Teda v prípade 95-percentného intervalu spoľahlivosti je $\alpha = 0{,}05$, a teda potrebujeme hodnotu $z_{0{,}975}$, pre ktorú je $\mathbb{P}(\mathcal{N}(0,1) \leq z_{0{,}975}) = 0{,}975$. Z obrázku 11 vidíme, že $z_{0{,}975} = 1{,}96$. 95-percentný interval spoľahlivosti teda je

$$\left[0{,}245 - 1{,}96 \cdot \sqrt{\frac{0{,}245 \cdot 0{,}755}{1000}}, 0{,}245 + 1{,}96 \cdot \sqrt{\frac{0{,}245 \cdot 0{,}755}{1000}}\right].$$

3 Stredná hodnota

Príklad 1. Výrobca batérií uvádza, že priemerná výdrž batérie ich nového modelu je 10 hodín. Kvôli opakovaným sťažnostiam na krátku výdrž batérií sa rozhodol preveriť toto tvrdenie a zistiť, či nie je priemerná výdrž menšia, ako by mala byť Z náhodnej vzorke 16 batérií boli zistené tieto charakteristiky o výdrži v hodinách (priemer a výberová disperzia):

$$\bar{x} = 9{,}4, \qquad s^2 = 1{,}2^2.$$

Predpokladajme, že výdrž batérií sa dá modelovať normálnym rozdelením.

(a) Sformulujte nulovú a alternatívnu hypotézu.

(b) Napíšte rozdelenie testovacej štatistiky za platnosti nulovej hypotézy. Vysvetlite základnú myšlienku, ako táto štatistika vznikla a prečo nemá normálne rozdelenie.

(c) Vypočítajte hodnotu testovacej štatistiky.

(d) Určte kritický obor pre test na hladine významnosti $\alpha = 0{,}05$. Použite pritom správnu hodnotu kvantilu z obrázku 12

(e) Rozhodnite, či nulovú hypotézu zamietame alebo nezamietame.

(f) Bolo povedané, že výdrž batérií sa dá modelovať normálnym rozdelením. Ako by sme toto tvrdenie otestovali, ak by sme mali k dispozícii všetkých 16 nameraných hodnôt (nielen priemer avýberovú disperziu)?

```python
from scipy.stats import t

df = 15
k = [0.005, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.995]
for i in k:
  print(f"P(t_{df} <= {t.ppf(i, df):.2f}) = {i}")
```

```
P(t_15 <= -2.95) = 0.005
P(t_15 <= -2.13) = 0.025
P(t_15 <= -1.75) = 0.05
P(t_15 <= -1.34) = 0.1
P(t_15 <= 1.34) = 0.9
P(t_15 <= 1.75) = 0.95
P(t_15 <= 2.13) = 0.975
P(t_15 <= 2.95) = 0.995
```

Obr. 12: Kvantily.

Riešenie.

(a) Označme μ skutočnú priemernú výdrž batérie. Chceme overiť, či výrobca výdrž batérií nepreceňuje, preto volíme jednostranný test:

$$H_0 : \mu = 10, \qquad H_1 : \mu < 10.$$

(b) Keďže nepoznáme disperziu výdrže batérií, použijeme jednovýberový t-test. Vo vzorcovníku nájdeme, že testovacia štatistika na testovanie hypotézy $\mu = \mu_0$ z n dát je

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}},$$

kde $\bar{X}$ je priemer a S^2 výberová dispezia. Táto štatistka má za platnosti nulovej hypotézy Studentovo t-rozdelenie $n - 1$ stupňami voľnosti.

Ak by sme poznali disperziu výdrže batérie σ^2, tak by za platnosti nulovej hypotézy platilo, že $\bar{X}$ má normálne rozdelenie so strednou hodnotou μ_0 a disperziou $\frac{\sigma^2}{n}$. V tejto situácii platí

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1).$$

V našom prípade disperziu nepoznáme a nahradíme ju výberovou disperziou. Pritom sa však normálne rozdelenie nezachová. Dá sa však odvodiť rozdelenie štatistiky, ktorá takto vznikne, a je to práve uvedené Studentovo t-rozdelenie.

(c) Hodnota testovacej štatistiky je:

$$T = \frac{9{,}4 - 10}{\sqrt{\frac{1{,}2^2}{16}}} = \frac{-0{,}6}{\frac{1{,}2}{4}} == \frac{-0{,}6}{0{,}3} = -2{,}0.$$

(d) Nulovú hypotézu v prospech alternatívnej zamietneme vtedy, ak bude priemerná výdrž batérií príliš nízka. V reči t-štatistiky to znamená, že hypotézu zamietneme pre príliš malú hodnotu štatistiky. Konkrétna hraničná hodnota sa učí tak, aby pravdepodobnosť zamietnutia pravdivej nulovej hypotézy bola 5 percent. Z obrázku 12 s kvantilmi vidíme, že musíme zobrať hodnotu -1,75 (ak platí nulová hypotéza, štatistika bude menšia ako -1,75 s pravdepodobnosťou práve 5 percent). Teda hypotézu zamietame, ak je štatistika menšia ako -1,75.

(e) Platí $-2 < -1{,}75$, a teda nulovú hypotézu zamietame. Tvrdíme teda na základe testov, že priemerná výdrž batérií nie je 10 hodín, ako sa to tvrdí, ale že je menšia.

(f) Kvôli malému počtu dát (16) je spomedzi testov, o ktorých sme hovorili, najvhodnejší Shapiro-Wilkov test.

Príklad 2. Výrobca balení ryže deklaruje, že priemerná hmotnosť jedného balenia je 1 kg. Z dôvodu technických obmedzení výrobnej linky môže dochádzať k miernemu odchýlkam a balenie môže mať o niečo väčšiu aj o niečo menšiu hmotnosť. Chceme však overiť, či nenastáva systematická chyba, teda či sa skutočná stredná hodnota líši od deklarovanej.

Zo vzorky 25 náhodne vybraných balení bol zistený priemer a výberová disperzia hmotnosti:

$$\bar{x} = 1{,}016, \qquad s^2 = 0{,}08^2.$$

Predpokladajme, že hmotnosť balení má približne normálne rozdelenie.

(a) Sformulujte nulovú a alternatívnu hypotézu.

(b) Vypočítajte hodnotu testovacej štatistiky a napíšte rozdelenie testovacej štatistiky za platnosti nulovej hypotézy.

(c) Určte zamietnutia pre test na hladine významnosti $\alpha = 0{,}05$. Použite pritom správnu hodnotu kvantilu z obrázku 13. V premennej `df` je vložený správny počet stupňov voľnosti.

(d) Rozhodnite, či nulovú hypotézu zamietame alebo nezamietame.

(e) Predpokladajme, že by sme spravili 95-percentný interval spoľahlivosti pre strednú hodnotu. Vieme bez jeho konkrétneho výpočtu povedať, či bude obsahovať hodnotu 1?

```python
from scipy.stats import t

k = [0.005, 0.025, 0.05, 0.1, 0.9, 0.95, 0.975, 0.995]
for i in k:
  print(f"P(t_{df} <= {t.ppf(i, df):.3f}) = {i}")
```

```
P(t_24 <= -2.797) = 0.005
P(t_24 <= -2.064) = 0.025
P(t_24 <= -1.711) = 0.05
P(t_24 <= -1.318) = 0.1
P(t_24 <= 1.318) = 0.9
P(t_24 <= 1.711) = 0.95
P(t_24 <= 2.064) = 0.975
P(t_24 <= 2.797) = 0.995
```

Obr. 13: Kvantily.

Riešenie.

(a) Označme μ skutočnú priemernú hmotnosť jedného balenia ryže. Zaujíma nás akákoľvek odchýlka od deklarovanej hodnoty, preto ide o obojstranný test:

$$H_0 : \mu = 1, \qquad H_1 : \mu \neq 1.$$

(b) Hodnota testovacej štatistiky je :

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} = \frac{1{,}016 - 1}{\sqrt{\frac{S^2}{n}}} = \frac{1{,}016 - 1}{\frac{0{,}08}{5}} = \frac{0{,}016}{0{,}016} = 1.$$

Za platnosti nulovej hypotézy má táto štatistika Studentovo t-rozdelenie s $n - 1$, v našom prípade 24 stupňami voľnosti.

(c) Nulovú hypotézu v prospech alternatívnej zamietneme vtedy, ak bude priemerná hmotnosť príliš nízka alebo príliš vysoká. V reči t-štatistiky to znamená, že hypotézu zamietneme pre príliš malú hodnotu štatistiky alebo príliš veľkú. Konkrétna hraničná hodnota sa učí tak, aby pravdepodobnosť zamietnutia pravdivej nulovej hypotézy bola 5 percent. Kvôli symetrii (nemáme dôvod preferovať niektorú možnosť) budeme chcieť, aby pravdepodobnosť zamietnutia kvôli príliš malej hodnote bola 2,5 percenta a kvôli príliš veľkej hodnote tiež 2,5 percenta.

Z obrázku 13 s kvantilmi vidíme, že musíme zobrať hodnoty -2,064 a 2,064. (ak platí nulová hypotéza, štatistika bude menšia ako -2,064 s pravdepodobnosťou práve 2,5 percenta a bude väčšia ako 2,064 tiež pravdepodobnosťou práve 2,5 percenta). Teda hypotézu zamietame, ak je štatistika menšia ako -2,064 alebo väčšia ako 2,064.

(d) Keďže $-2{,}064 < 1 < 2{,}064$, nulovú hypotézu o strednej hodnote 1 kg nezamietame.

(e) To, že sa nejaká hodnota m nachádza v 95-percentnom intervale spoľahlivosti je ekvivalentné s tým, že sa pri testovaní na 5-percentnej hladine významnosti nezamietne hypotéza $\mu = m$. V našom prípade sa hypotéza o strednej hodnote 1 nezamietla, a teda 95-percentný interval spoľahlivosti hodnotu 1 obsahuje.

Príklad 3. Uvádza sa, že priemerná rozpustnosť šumivej tablety v pohári vody je dva a pol minúty. Janko Hraško si zapisoval časy, ako dlho sa rozpúšťali tablety v jeho balení. Výsledok testovania je na obrázku 14. Predpokladjme, že čas, za ktorý sa tableta rozpustí, má normálne rozdelenie.

```
from scipy import stats

stats.ttest_1samp(data, popmean=2.5)
```

```
TtestResult(statistic=np.float64(-2.0630658981652052),
pvalue=np.float64(0.046809680611503826), df=np.int64(34))
```

Obr. 14: Výpočet k príkladu 4.

(a) Aká hypotéza sa testovala a s akou alternatívnou hypotézou?

(b) Aký je výsledok testu? Uvažuje hladiny významosti 1 percento, 5 percent, 10 percent.

(c) Pomocou koľkých dát bola robená táto analýza?

Riešenie.

(a) Testovala sa hypotéza o strednej hodnote času potrebného na rozpustenie tablety. Alternatívna hypotéza nebola špecifikovaná, takže sa použila defaultná obostranná. Teda

$$H_0 : \mu = 2{,}5, \qquad H_1 : \mu \neq 2{,}5.$$

(b) P-hodnota vo výstupe je 0,0468. Nulovú hypotézu zamietneme, ak je p-hodnota menšia než zvolená hladina významnosti. Pre hladinu významnosti 0,01 (1 percento) je p-hodnota väčšia ako táto hladina významnosti, takže na tejto hladine hypotézu nezamietame. Pri hladine významnosti 0,05 (5 percent) je p-hodnota menšia ako táto hladina významnosti a nulovú hypotézu zamietame. Rovnaký záver platí aj pri hladine významnosti 0,1, (10 percent) kde je p-hodnota znovu menšia ako hladina významnosti, a preto aj tu nulovú hypotézu zamietame.

(c) Vo výstupe vidíme počet stupňov voľnosti 34, pričom vieme, že sa rovná $n-1$, kde n je počet dát. Teda v tomto prípade bolo použitých 35 dát.

Príklad 4. Skupina študentov dostala na začiatku vyučovacej hodiny z angličtiny krátky test z predbranej lekcie. Počas hodiny sa lekcia zopakovala a na jej konci študenti znovu napísali rovnaký test.

Na obrázku 15 je výstup zo štatistického testu. V premenných `pred` a `po` sú zaznamenané body, ktoré študenti získali pred opakovaním lekcie a po jeho opakovaní.

```
stats.ttest_rel(pred, po, alternative="less")
```

```
TtestResult(statistic=np.float64(-2.098631773120034),
pvalue=np.float64(0.023514631133244878), df=np.int64(23))
```

Obr. 15: Výstup zo štatistického testu.

(a) Aký test sa použil na analýzu výsledkov? Prečo je vhodný?

(b) Sformulujte nulovú a alternatívnu hypotézu, ktorá bola použitá. Prečo je v tomto prípade rozumné použiť takúto alternatívu?

(c) Aký predpoklad o dátach sa robí pri použití tohto testu?

(d) Aký je záver testu na 5-percentnej hladine významnosti?

(e) Ako by sme mohli ekvivalentne postupovať pomocou jednovýberového t-testu?

Riešenie.

(a) Pre analýzu výsledkov bol použitý párový t-test (funkcia `ttest_rel`). Tento test je vhodný, pretože ide o dve merania na tej istej skupine študentov pred a po opakovaní lekcie, pričom chceme zistiť, či sa priemerný počet bodov skóre. Párový t-test berie do úvahy závislosť dát - pre každého študenta sa počíta jeho rozdiel pre a po opakovaní.

(b) Parameter `alternative` je zvolený ako `less`, čo znamená, že stredná hodnota prvej premennej je menšia ako stredná hodnota druhej premennej. Nulová a alternatívna hypotéza teda sú:

$$H_0 : \mu_{\text{po}} = \mu_{\text{pred}}, \quad \text{teda že opakovanie lekcie nemá vplyv na body z testu,}$$

$$H_1 : \mu_{\text{pred}} < \mu_{\text{po}}, \quad \text{teda že opakovanie lekcie zlepšuje výsledok testu.}$$

Je rozumné použiť jednostrannú alternatívu, pretože očakávame, že opakovanie lekcie zlepšuje výsledky. Pri obojstrannej alternatíve by sme po zamietnutí nulovej hypotézy mohli povedať iba to, že výsledky pre a po opakovaní nie sú rovnaké, ale nemohli by sme povedať, že opakovanie výsledky zlepšilo.

(c) Predpokladáme, že rozdiely pred a po opakovaní, vypočítané pre každého študenta, majú normálne rozdelenie.

(d) P-hodnota je 0,0235, čo je menej ako 0,05. Na 5-percentnej hladine významnosti teda nulovú hypotézu zamietame. Tvrdíme teda, že opakovanie zlepšuje výsledok jazykového testu.

(e) Mohli by sme vytvoriť dáta s rozdielmi, napríklad ako `rozdiely = [a-b for a,b in zip(pred, po)]` a potom pomocou jednovýberového testu testovať, či majú nulovú strednú hodnotu s alternatívou, že stredná hodnota je záporná.

Príklad 4. Na základnej škole chcú zistiť, či výučba matematiky podľa novej učebnice ovplyvní výsledky žiakov. V ročníku sú dve triedy. Trieda A absolvovala tradičnú výučbu, v triede B používali novú učebnicu. Na konci školského roka obidve triedy dostali rovnakú písomku. Úlohou je zistiť, či sa priemerný počet bodov žiakov v triede A líši od priemeru v triede B.

```
stats.ttest_ind(trieda_A, trieda_B)
```

```
TtestResult(statistic=np.float64(-1.3976610866242438),
pvalue=np.float64(0.16973201348509), df=np.float64(41.0))
```

Obr. 16: Výstup zo štatistického testu.

(a) Aký test sa použil na analýzu výsledkov? Prečo je vhodný?

(b) Sformulujte nulovú a alternatívnu hypotézu, ktorá bola použitá. Prečo bola takto zvolená alternatíva?

(c) Aký predpoklad o dátach sa robí pri použití tohto testu?

(d) Aký je záver testu na 5-percentnej hladine významnosti?

(e) Je aj v tomto prípade možné ekvivalentne postupovať pomocou jednovýberového t-testu (tak ako v predchádzajúcom prípade)?

Riešenie.

(a) Použil sa nepárový t-test (funkcia `ttest_ind`). Dôvod je ten, že skupiny A a B sú nezávislé – žiaci z jednej skupiny nie sú párovaní so žiakmi z druhej skupiny, takže párový test sa nedá použiť.

(b) Parameter `alternative` nebol špecifikovaný, defaultná hodnota je obojstranná alternatíva, čo znamená, že stredná hodnota prvej premennej je iná ako stredná hodnota druhej premennej. Nulová a alternatívna hypotéza teda sú:

$$H_0 : \mu_A = \mu_B, \quad \text{teda že stredné hodnoty v oboch triedach sú rovnaké,}$$

$$H_1 : \mu_A \neq \mu_B, \quad \text{teda že stredné hodnoty sú rôzne}$$

Takáto voľba alternatívnej hypotézy znamená, že nevieme dopredu povedať, či nový spôsob výučby zlepší alebo zhorší výsledky. Preto testujeme akékoľvek odchýlky od skupiny s tradičnou výučbou.

(c) Predpokladáme normálne rozdelenie počtu bodov v jednotlivých triedach. Okrem toho, keďže nebol špecifikovaný parameter `equal_var`, použila sa jeho defaultná hodnota `True` - to znamená, že predpokladáme rovnakú disperziu v obidvoch triedach. Obidva predpoklady sa dajú testovať pomocou štatistických testov.

(d) P-hodnota vo výstupe vyšla 0,1697, čo je viac ako 0,05. Na 5-percentnej hladine významnosti teda nulovú hypotézu nezamietame. Tvrdíme teda, že stredná hodnota počtu bodov je rovnaká pri obidvoch spôsoboch výučby.

(e) Jednovýberový t-test sa pri takýchto dátach použiť nedá.

Príklad 5. Na konci školského roka absolvovali žiaci štyroch tried záverečný test z matematiky. Cieľom je zistiť, či sa priemerné výsledky žiakov jednotlivých tried líšia.

(a) Ktorý štatistický test je na obrázku 17 použitý na porovnanie priemerov štyroch tried? Sformulujte nulovú a alternatívnu hypotézu pre tento test. Aké sú predpoklady použitia tohto testu?

```
stats.f_oneway(A, B, C, D)
```

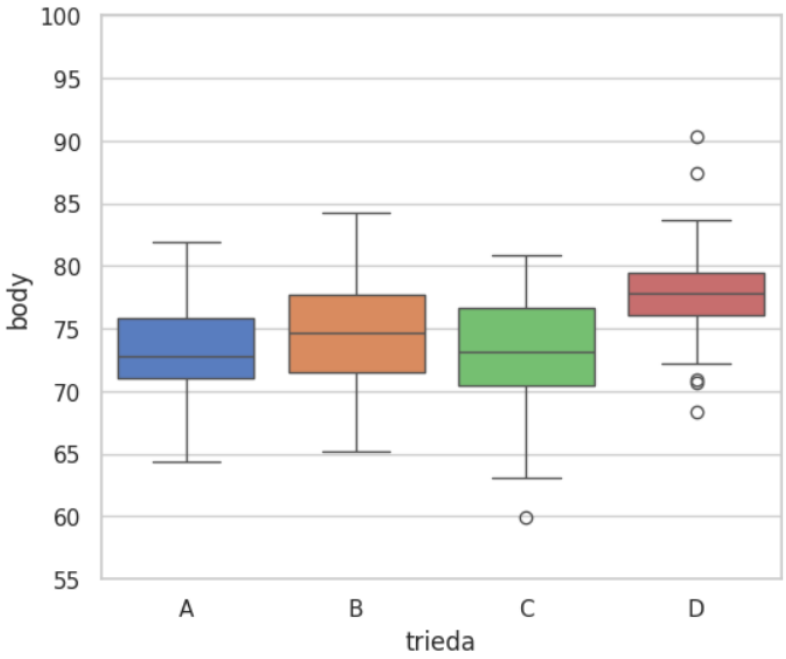
```
F_onewayResult(statistic=np.float64(7.183741497286111),
pvalue=np.float64(0.00018226528812997176))
```

Obr. 17: Výstup zo štatistického testu.

(b) Na základe výstupu rozhodnite, či na hladine významnosti 5 percent zamietneme nulovú hypotézu alebo nie.

(c) Navrhnite postup, ako zistiť, medzi ktorými triedami sú rozdiely.

(d) Na obrázku 18 je zobrazený boxplot bodov podľa tried. Popíšte, čo na základe týchto boxplotov intuitívne očakávame, že výjde v predchádzajúcom bode.



Obr. 18: Boxploty bodov podľa tried.

Príklad 5. Na konci školského roka absolvovali žiaci štyroch tried záverečný test z matematiky. Cieľom je zistiť, či sa priemerné výsledky žiakov jednotlivých tried líšia.

(a) Použila sa ANOVA (funkcia `F_oneway`). Hypotézy sú

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D,$$

teda že stredné hodnoty vo všetkých triedach sú rovnaké, Alternatívna hypotéza je, že stredné hodnoty nie sú všetky rovnaké (negácia nulovej hypotézy). Predpokladáme normálne rozdelenie počtu bodov v jednotlivých triedach. Okrem toho, keďže nebol špecifikovaný parameter `equal_var`, použila sa jeho defaultná hodnota `True` - to znamená, že predpokladáme rovnakú disperziu v obidvoch triedach. Obidva predpoklady sa dajú testovať pomocou štatistických testov.

(b) P-hodnota vo výstupe vyšla 0,000182, čo je menej ako 0,05. Na 5-percentnej hladine významnosti teda nulovú hypotézu zamietame. Tvrdíme teda, že stredná hodnota počtu bodov nie je vo všetkých triedach rovnaká.

(c) Na zistenie, medzi ktorými triedami sú rozdiely, budeme testovať zhody stredných hodnôt pre všetky dvojice tried. Na to môžeme použiť nepárový t-test.

Treba ešte myslieť na to, že budeme robiť 6 porovnaní. Pri takomto viacnásobnom testovaní sa zvyšuje riziko falošne pozitívnych výsledkov, teda že niektoré hypotézy zamietneme „zbytočne“. Pri viacerých testoch už neplatí, že pravdepodobnosť zamietnutia pravdivej nulovej hypotézy je 5 percent. To platí pre každý jeden test samostatne, ale ak ich je viac, pravdepodobnosť takejto chyby

pri niektorom teste je vyššia. Aby sme tomu predišli, používajú sa úpravy p-hodnôt.

(d) Z grafu sa zdá, že v triede D je stredná hodnota iná ako v ostatných.

4 Čo ešte zostalo...

Príklad 1. Na obrázku 19 je najskôr použitý genrátor náhodných čísel, ktorý funguje tak, že vrátu číslo, ktoré vznikne ako súčet 12 nezávislých náhodných čísel s rovnomerným rozdelením na $(0, 1)$, od ktorého sa odčíta 6.

```
def generator():
    return np.sum(np.random.uniform(0, 1, 12)) - 6

n = 10**4
data = [generator() for _ in range(n)]

stats.kstest(data, 'norm', args=(0,1))
```

```
KstestResult(statistic=np.float64(0.010120374563509449),
pvalue=np.float64(0.25559982666465286),
statistic_location=np.float64(-0.6495866621536699),
statistic_sign=np.int8(1))
```

Obr. 19: Kód k príkladu 1.

(a) Aký test sa použil na takto vygenerované dáta? Aká hypotéza sa testovala a s akým výsledkom?

(b) Ako sa dá vysvetliť výsledok tohto testu?

Riešenie

(a) Použil sa Kolmogorovov-Smirnovov test, ktorým sa testovala zhoda s normálnym rozdelením s parametrami 0 a 1 (teda zhoda s normalizovaným normálnym rozdelením). P-hodnota vo výstupe vyšla 0,2566, čo je viac ako 0,05. Na 5-percentnej hladine významnosti teda nulovú hypotézu o normalizovanom normálnom rozdelení nezamietame.

(b) Takto generované náhodné čísla normálne rozdelenie nemajú. Dá sa to vidieť z toho, že náhodná premenná s normálnym rozdelením môže nadobúdať ľubovoľné reálne hodnoty. Nami generované náhodné čísla túto vlastnosť nemajú. Každý z dvanástich sčítancov nadobúda hodnoty z intervalu $(0, 1)$, ich súčet je teda z intervalu $(0, 12)$. Po odpočítaní hodnoty 6 dostaneme číslo z intervalu $(-6, 6)$.

Dôvodom, prečo sa rozdelenie generovaných náhodných čísel podobá na normálne, je centrálna limitná veta. Podľa nej má súčet nezávislých rovnako rozdelených náhodných premenných približne normálne rozdelenie. So zväčšujúcim sa počtom sčítancov sa k normálnemu rozdeleniu približuje, pričom vidíme, že v tomto prípade stačí na veľmi dobrú zhodu už 12 sčítancov (keďže po odpočítaní konštanty 6 sa zhoda s normálnym rozdelením zachová). Stredná hodnota a disperzia je rovnaká ako pre testované normálne rozdelenie. Zo vzorcovníka z pravdepodobnosti vieme, že rovnomerné rozdelenie na intervale $(0, 1)$

má strednú hodnotu $\frac{1}{2}$ a disperziu $\frac{1}{12}$. Ak teda označíme $X_1, \ldots, X_{12}$ nezávislé náhodné premenné s rovnomerným rozdelením na $(0, 1)$, tak

$$\mathbb{E}(X_1 + \ldots X_{12} - 6) = \mathbb{E}(X_1) + \ldots \mathbb{E}(X_{12}) - 6 = 12 \cdot \frac{1}{2} - 6 = 0,$$

$$\mathbb{D}(X_1 + \ldots X_{12} - 6) = \mathbb{D}(X_1 + \ldots X_{12}) = \mathbb{D}(X_1) + \ldots \mathbb{D}(X_{12}) - = 12 \cdot \frac{1}{12} = 1.$$

Príklad 2. Vo viacerých príkladoch v predchádzajúcej kapitole sme potrebovali, aby boli dáta normálne rozdelené. Mohli by sme použiť Kolmogorovov-Smirnovov test z predchádzajúceho príkladu na otestovanie normálneho rozdelenia týchto dát?

Riešenie Nie, Kolmogorovov-Smirnovov test sa v takýchto príkladoch nedá použiť. Tento test predpokladá, že parametre testovaného rozdelenia sú dané, teda že nie sú odhadované z dát. To nie je prípad testovania normálneho rozdelenia pre zadané dát - vtedy parametre nepoznáme.

Dá sa namiesto toho použiť Lillieforsov test, ktorý používa rovnakú myšlinenku ako Kolmogorovov-Smirnovov test - porovnanie embirickej distribučnej funkcie z dát a distribučnej funkcie testovaného rozdelenia. Sú aj iné testy - spomínali sme D'Agostinov-Pearsonov test a Shapirov-Wilkov test (ten sa dá použiť aj v prípade, keď máme malý počet dát).

Vzorcovník II. na skúšku (štatistika)

- **Chí-kvadrát test dobrej zhody**
  - Testujeme zhodu dát s daným rozdelením
  - Testovacia štatistika: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$, kde O_i sú pozorované počty (observed) a E_i sú očakávané počty (expected)
  - V Pythone funkcia `chisquare(f_obs, f_exp)`
  - Za platnosti nulovej hypotézy sa rozdelenie štatistiky dá aproximovať chí-kvadrát rozdelením s $k-1$ stupňami voľnosti.
  - Ak sa odhaduje p parametrov, počet stupňov voľnosti sa zmenší ešte o p (v Pythone parameter `ddof`)
- **Iné testy zhody dát so zadaným rozdelením**
  - Kolmogorovov-Smirnovov test: zhoda so spojitým rozdelením, ktorého parametre nie sú odhadované (v Pythone `kstest`)
  - Lillieforsov test - modifikácia Kolmogorovovho-Smirnovovho testu pre testovanie normálneho rozdelenia s neznámymi parametrami
  - Ďalšie testy normálneho rozdelenia: D'Agostinov-Pearsonov test (v Pythone `normaltest`) a Shapirov-Wilkov test (v Pythone `shapiro`, dá sa použiť aj v prípade, keď máme malý počet dát).
- **Test o pravdepodobnosti úspechu v postupnosti nezávislých pokusov**
  - Testuje sa $p = p_0$
  - Testovacia štatistika: $Z = \frac{\bar{X} - np_0}{\sqrt{np_0(1-p_0)}}$, kde X je počet úspechov v n pokusoch
  - Za platnosti nulovej hypotézy má približne $\mathcal{N}(0,1)$ rozdelenie
  - V Pythone: `proportions_ztest(count, nobs)`, parameter `alternative` je defaultne `two-sided`, nastavíme ešte `prop_var=False`
- **Test o strednej hodnote normálneho rozdelenia**
  - Disperzia normálneho rozdelenia je neznáma
  - Testuje sa $\mu = \mu_0$, alternatíva môže byť $\mu \neq \mu_0$, $\mu > \mu_0$, $\mu < \mu_0$
  - V Pythone `ttest_1sample`, parameter `popmean` je μ_0, alternatíva je defaultne `two-sided`, nerovnosti sa zapíšu ako `less`, `greater`
  - Testovacia štatistika: $T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}}$
  - Za platnosti nulovej hypotézy má Studentovo t-rozdelenie $n-1$ stupňami voľnosti.
- **Párový t-test**

  - párové dáta,
  - vyžaduje normálne rozdelenie
  - testuje sa $\mu_1 = \mu_2$
  - v Pythone `ttest_rel`
  - parameter `alternative` môže byť `two-sided` (default), `less`, `greater`

- **Nepárový t-test**
  - nepárové dáta,
  - vyžaduje normálne rozdelenie
  - testuje sa $\mu_1 = \mu_2$
  - v Pythone `ttest_ind`
  - parameter `alternative` môže byť `two-sided` (default), `less`, `greater`
  - parameter `equal_var` - či sú rovnaké disperzie, defaultne `True`

- **ANOVA**
  - vyžaduje normálne rozdelenie
  - testuje sa zhoda všetkých stredných hodnôt (alternatíva, že nie sú rovnaké)
  - v Pythone `F_oneway`
  - parameter `equal_var` - či sú rovnaké disperzie, defaultne `True`

- **Intervaly spoľahlivosti**
  - $100(1-\alpha)$-percentný IS pre pravdepodobnosť úspechu v pokuse:

$$\left[\hat{p} - z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

    kde z_q je také číslo, ktoré spĺňa podmienku $\mathbb{P}(\mathcal{N}(0,1) \leq z_q) = q$
  - $100(1-\alpha)$-percentný IS pre strednú hodnotu normálneho rozdelenia:

$$\left[\bar{X} - t_{1-\frac{\alpha}{2},n-1}\sqrt{\frac{S^2}{n}},\ \bar{X} + t_{1-\frac{\alpha}{2},n-1}\sqrt{\frac{S^2}{n}}\right]$$

    kde $t_{q,k}$ je také číslo, ktoré spĺňa podmienku $\mathbb{P}(X \leq t_{q,k}) = q$ pre X so Studentovým rozdelením s k stupňami voľnosti

- **Pravdepodobnostné rozdelenia v Pythone**: `pmf` - pravdepodobnosť (probability mass function), `cdf` - distribučná funkcia (cummulative distribition function), `ppf` - inverzná funkcia k distribučnej (percent point function)