DÚ č.2 - Testování hypotéz

Marek Nevole, Jan Novotný

ČVUT - FIT

{nevolmar, novot103}@fit.cvut.cz

13. března 2022

1 Úvod

Ve druhém úkolu z předmětu vybrané statistické metody jsme se zabývali testováním hypotéz a testy nezávislosti. Za reprezentanta byl zvolen Marek Nevole.

$$K = 28$$
$$L = 6$$
$$X = ((23KL) \bmod 20) + 1$$
$$X = 5$$
$$Y = ((X + ((5K + 7L) \bmod 19)) \bmod 20) + 1$$
$$Y = 17$$

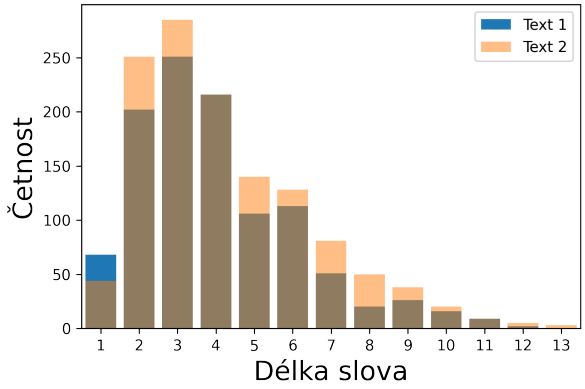
Výsledkem těchto rovnic jsou názvy vybraných datových souborů. V našem případě budeme pracovat se soubory 005.txt a 017.txt.

Úkol jsme vypracovali pomocí programovacího jazyka Python[1] v prostředí Jupyter Notebook[2] s volně dostupnými knihovnami SciPy[3], NumPy[4] a Matplotlib[5].

2 Úloha č.1

Z obou datových souborů načtěte texty k analýze. Pro každý text zvlášť odhadněte základní charakteristiky délek slov, tj. střední hodnotu a rozptyl. Graficky znázorněte rozdělení délek slov.

Texty datových souborů jsme upravili tak, aby odpovídaly náhodným veličinám délek slov. Základní charakteristiky jsme odhadli pomocí bodových odhadů, konkrétně výběrovým průměrem a výběrovým rozptylem s jmenovatelem $n - 1$, který dělá rozptyl nestranný. K výpočtu jsme použili knihovnu NumPy a její funkce *mean* a *var* s parametrem *ddof=1*, který zajistí požadovanou nestrannost rozptylu. Náhodnou veličinu délky slov textu 005 jsme označili jako X a podobně pro text 017 Y. Grafické



Obrázek 1: Rozdělení délek slov obou textů.

znázornění rozdělení délek slov lze pozorovat na obrázku 1, který jsme vytvořili jako barový graf, který připomíná histogram, kde pro každou hodnotu je vytvořen samostatný bin.

$$\bar{X}_n = 4.010, s_X^2 = 4.451$$
$$\bar{Y}_n = 4.283, s_Y^2 = 5.073$$

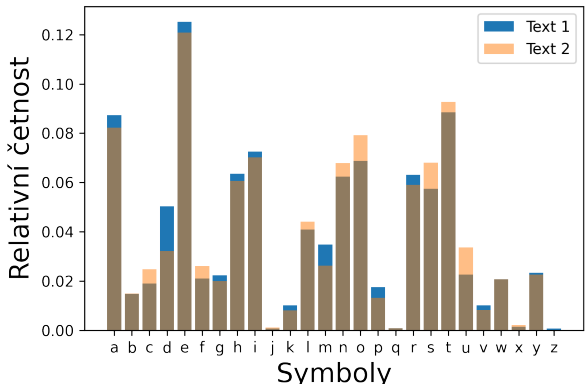
```
import numpy as np

l1 = [len(w) for w in txt1.split()]
sample_mean_1 = np.mean(l1)
sample_var_1 = np.var(l1, ddof = 1)
```

3 Úloha č.2

Pro každý text zvlášť odhadněte pravděpodobnosti písmen (symbolů mimo mezery), které se v textech vyskytují. Výsledné pravděpodobnosti graficky znázorněte.

Pro úlohu č.2 jsme přemodelovali náhodné veličiny na četnosti znaků, což svým rozdělením odpovídá multinomickému rozdělení. Pravděpodobnosti písmen jsme odhadli jako relativní četnosti, tedy počet výskytů daného symbolu vydělený počtem všech

[1] python.org
[2] jupyter.org
[3] scipy.org
[4] numpy.org
[5] matplotlib.org



Obrázek 2: Relativní četnosti písmen obou textů.

symbolů. Tyto pravděpodobnosti jsou znázorněny na obrázku 2.

4 Úloha č.3

Na hladině významnosti 5% otestujte hypotézu, že rozdělení délek slov nezávisí na tom, o který jde text. Určete také p-hodnotu testu.

Cílem třetí úlohy bylo rozhodnout zda rozdělení délek slov nezávisí na tom, o který jde text. Tedy jsme testovali, nezávislost mezi veličinami náhodného vektoru $(X, Y)^T$, kde náhodná veličina X určuje číslo textu a Y délku slova. Tento vektor má diskrétní rozdělení. Na základě dat jsme sestavili kontingenční tabulku. Za nulovou hypotézu H_0 jsme postavili nezávislost těchto veličin, tedy $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$, oproti alternativní hypotéze $H_A : p_{ij} \neq p_{i\bullet}p_{\bullet j}$. Test nezávistlosti pomocí Pearsonovy statistiky χ^2 jsme nemuseli dělat ručně, jelikož je součástí knihovny Scipy jako funkce *chi2_contingency*. Předtím než jsme provedli samotný test, bylo důležité zkontrolovat teoretické četnosti vypočítané z kontingenční tabulky pomocí funkce *expected_freq* ze stejné knihonvy. Testová statistika χ^2 je pouze asymptotická, a při nizkých teoretických četnostech $np_{ij} \leq 5$ dává nepřesné výsledky. Vypočítané teoretické četnosti byly nedostačující u nejdelších slov. Jako řešení jsme tato slova počítali jako o písmeno kratší slova, dokud jsme nedosáhli požadovaných četností. Poté už jsme mohli použít zmíněnou funkci. Výstup funkce je testová statistika, p-hodnota, stupně volnosti a teoretické četnosti. Podle $\hat{p} = 0.002$ jsme zamítli nulovou hypotézu H_0 ve prospěch alternativní hypotézy H_A. Pro jistotu jsme také spočítali kritickou hodnotu $\chi^2_{\alpha,(r-1)(c-1)}$, kde $\alpha = 0.05, r = 11$ a $c = 2$. Dle očekávání platí $\chi^2 \geq \chi^2_{0.05,10}$.

$$\chi^2 = 26.7$$
$$\hat{p} = 0.002$$
$$\chi^2_{0.05,10} = 18.307$$

5 Úloha č.4

Na hladině významnosti 5% otestujte hypotézu, že se střední délky slov v obou textech rovnají. Určete také p-hodnotu testu.

Ve čtvrté úloze jsme testovali hypotézu, zda se střední délky slov v obou textech rovnají. Jako nulovou hypotézu H_0 jsme postavili $\mu_1 = \mu_2$ oproti alternativní $H_A : \mu_1 \neq \mu_2$.

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Pro tento týp úlohy je vhodný dvouvýběrový t-test. Předtím než tento test můžeme využít musíme zjistit, zda se rozptyly se rovnají nebo nerovnají. K tomu lze využít jiný test. Na základě nejistoty v normalitu X a Y jsme se rozhodli pro využití Levenova testu rovnosti rozptylů σ_1^2, σ_2^2, který je v tomto případě výhodnější. Za hladinu významnosti volíme 5%.

$$W = 4.838$$
$$\hat{p} = 0.0279$$

Dle p-hodnoty $\hat{p}$, která je menší než zvolená hodnota α, můžeme zamítnout hypotézu H_0, ve prospěch H_A, podle které se hodnoty rozptylů nerovnají. Toto můžeme ověřit nahlédnutím do úlohy 1, ve které jsme bodově odhadovali tyto rozptyly, kde se rozptyly opravdu nepodobají. Na základě předchozího výsledku jsme již mohli využít spravný typ dvouvýběrového t-testu s rozdílnými rozptyly. V knihovně SciPy je pro tento test funkce *ttest_ind*, která nabízí parametr rovnosti rozptylů *equal_var*.

$$T = -3.033$$
$$\hat{p} = 0.00244$$

Na základě $\hat{p}$ jsme zamítli H_0 ve prospěch H_A. Tedy střední délky slov obou textů se nerovnají.

6 Úloha č.5

Na hladině významnosti 5% otestujte hypotézu, že rozdělení písmen nezávisí na tom, o který jde text. Určete také p-hodnotu testu.

Tabulka 1: Kontingenční tabulka třetí úlohy.

	1	2	3	4	5	6	7	8	9	10	11+	$\sum$
005	68	202	251	216	106	113	51	20	26	16	11	1080
017	44	251	285	216	140	128	81	50	38	20	17	1270
$\sum$	112	453	536	432	246	241	132	70	64	36	28	2350

Podobně jako ve třetí úloze jsme v této úloze testovali nezávislost dvou veličin náhodného vektoru $(X, Y)^T$, kde náhodná veličina X určuje číslo textu, ale Y již neurčuje délku slov, ale četnosti symbolů. Postupovali jsme velice podobně. Za nulovou hypotézu H_0 jsme postavili nezávislost těchto veličin, tedy $H_0 : p_{ij} = p_{i\bullet}p_{\bullet j}$, oproti alternativní hypotéze $H_A : p_{ij} \neq p_{i\bullet}p_{\bullet j}$. Dále jsme na základě teoretických četností sestavili kontingenční tabulku tak, abychom mohli použít asymptotickou χ^2 statistiku. Kvůli nedostatku četností pro písmena *j, q, x, z*, jsme tyto písmena spojili do jednoho symbolu. Poté jsme opět využili funkci *chi2_contingency*.

$$\begin{aligned} \chi^2 &= 60.431 \\ \hat{p} &= 1.9 \times 10^{-5} \\ \chi^2_{0.05,22} &= 33.92 \end{aligned}$$

Dle p-hodnoty jsme zamítli nulovou hypotézu ve prospěch té alternativní, říkající, že rozdělení písmen závisí na daném textu.