

MI–SPI Druhá úloha

Tomáš Pšenička, Jan Groschaft

22. května 2020



Obsah

1	Zadání	2
2	Řešení	2
2.1	Charakteristika délek slov	2
2.2	Pravděpodobnosti znaků	3
2.3	Na hladině významnosti 5% otestujte hypotézu, že rozdělení délek slov nezávisí na tom, o který jde text. Určete také p-hodnotu testu.	3
2.4	Na hladině významnosti 5% otestujte hypotézu, že se střední délky slov v obou textech rovnají. Určete také p-hodnotu testu.	5
3	Závěr	6

1 Zadání

- Z obou datových souborů načtete texty k analýze. Pro každý text zvlášť odhadněte základní charakteristiky délek slov, tj. střední hodnotu a rozptyl. Graficky znázorněte rozdělení délek slov.
- Pro každý text zvlášť odhadněte pravděpodobnosti písmen (symbolů mimo mezery), které se v textech vyskytují. Výsledné pravděpodobnosti graficky znázorněte.
- Na hladině významnosti 5% otestujte hypotézu, že rozdělení délek slov nezávisí na tom, o který jde text. Určete také p-hodnotu testu.
- Na hladině významnosti 5% otestujte hypotézu, že se střední délky slov v obou textech rovnají. Určete také p-hodnotu testu.
- Na hladině významnosti 5% otestujte hypotézu, že rozdělení písmen nezávisí na tom, o který jde text. Určete také p-hodnotu testu.

2 Řešení

2.1 Charakteristika délek slov

Pro odhad charakteristik délek slov jsme použili výběrový průměr a výběrový rozptyl:

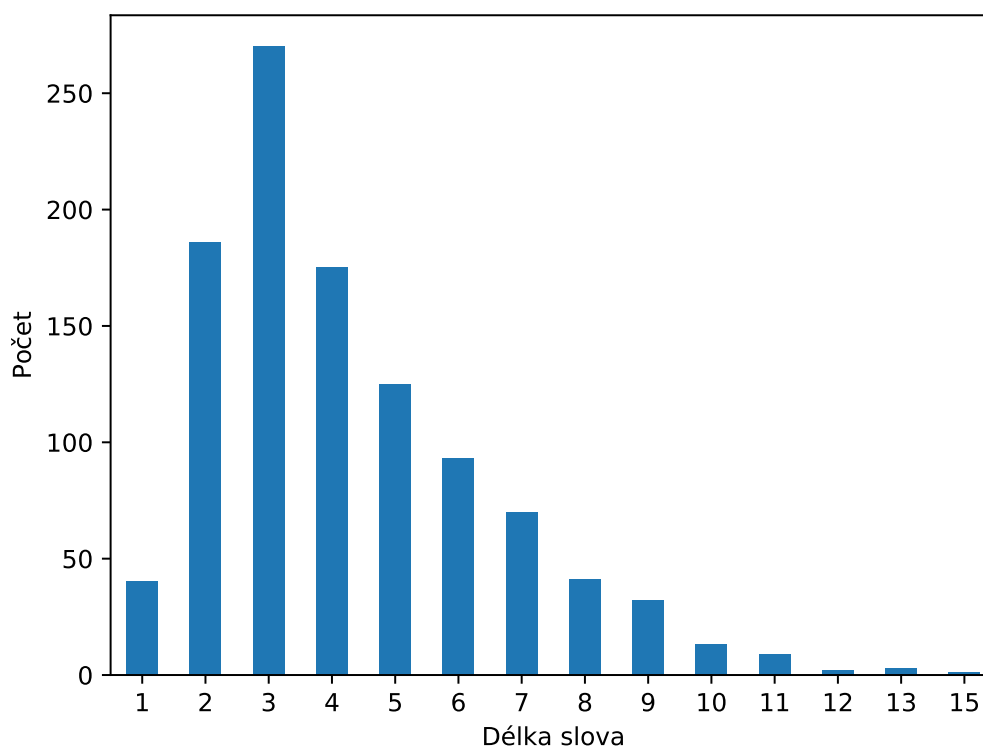
$$\bar{X}_n = \frac{1}{n} \sum_i X_i$$

$$s_n^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X}_n)^2$$

. Výsledky sledujeme v tabulce 1 a na obrázcích 1 a 2.

Text	\bar{X}_n	s_n^2
009.txt	4.2623	4.9860
011.txt	3.8603	4.5360

Tabulka 1: Charakteristika délek slov



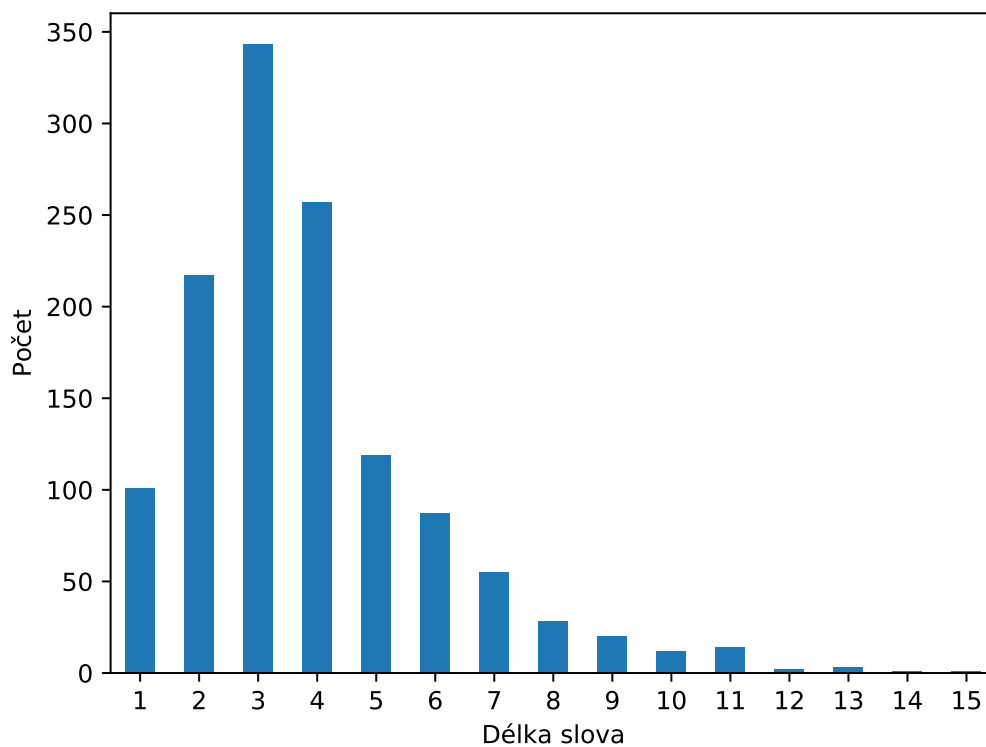
Obrázek 1: Rozdělení délek slov v souboru 009.txt

2.2 Pravděpodobnosti znaků

Pravděpodobnost jednotlivých znaků byla vypočítána jako poměr počtu výskytů jednotlivých znaků k celkovému počtu všech znaků a je znázorněna v grafech 3 a 4 pomocí nástroje matplotlib.

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

Provedli jsme test nezávislosti v kontingenční tabulce, viz tabulku 2. Testujeme hypotézu H_0 , že rozdělení délek slov nezávisí na volbě zdrojového textu oproti alternativě H_A , že na volbě zdrojového textu závisí.



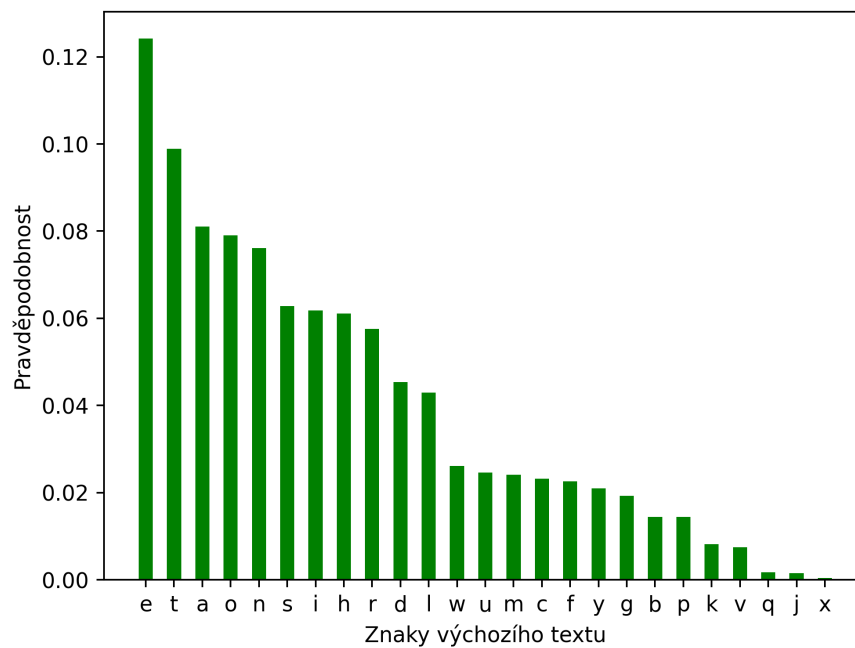
Obrázek 2: Rozdělení délek slov v souboru 011.txt

Testová statistika χ^2 má tvar

$$\chi^2 = n \sum_{i=1}^r \sum_{j=1}^c \frac{N_{i,j}^2}{N_{i,\bullet} N_{\bullet,j}}$$

a po dosazení $r = 2, c = 15$ máme $\chi^2 = 45.6235$ a p-hodnotu $3.2990e - 05$

Kritická hodnota je $\chi_{\alpha,(r-1)(c-1)}^2 = \chi_{0.05,14}^2 = 23.68$. Protože je naše testová statistika větší než kritická hodnota, zamítáme nulovou hypotézu ve prospěch alternativy.



Obrázek 3: Grafické znázornění pravděpodobností znaků textu 009.txt

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

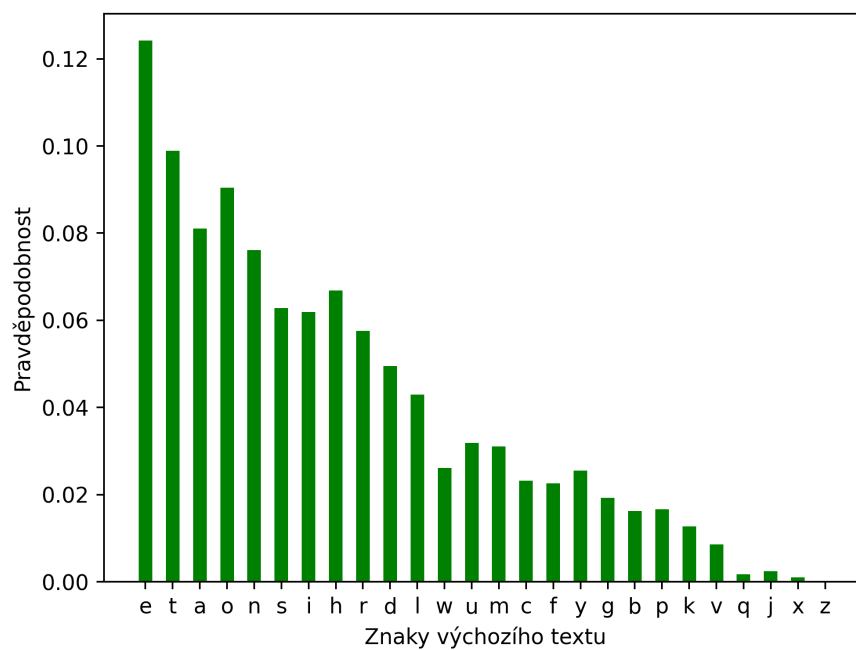
Chceme testovat hypotézu $H_0 : \mu_1 = \mu_2$ proti $H_A : \mu_1 \neq \mu_2$. Použijeme dvouvýběrový t-test za předpokladu různých rozptylů, tedy $\sigma_1 \neq \sigma_2$.

Použijeme tedy testovou statistiku

$$T = \frac{\bar{X}_n - \bar{Y}_n}{s_d},$$

kde

$$s_d = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$$



Obrázek 4: Grafické znázornění pravděpodobností znaků textu 011.txt

Soubor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
009.txt	40	186	270	175	125	93	70	41	32	13	9	2	3	0	1
011.txt	101	217	343	257	119	87	55	28	20	12	14	2	3	1	1

Tabulka 2: Kontingenční tabulka četností délek slov

3 Závěr