

Kritická práce s daty

4

Radek Čech

Statistické testy významnosti

- porovnávají se dvě hypotézy
 - **nulová hypotéza:**
tvrzení, které obvykle deklaruje “žádný rozdíl”, tj. nalezený rozdíl je dán variabilitou dat, náhodou (např. mince není falešná; mezi formou jazyka a četností užívání *bychom/bysme* není rozdíl)
 - **alternativní hypotéza:**
situace, kdy nulová hypotéza neplatí, tj. mezi proměnnými se předpokládá závislost; důležité je přitom nějaké teoretické zdůvodnění

Kardinální proměnné

- nabývají číselných hodnot

Kardinální proměnné

- nabývají číselných hodnot
 - délka slov, vět

Kardinální proměnné

- nabývají číselných hodnot
 - délka slov, vět
 - trvání slabik

Kardinální proměnné

- nabývají číselných hodnot
 - délka slov, vět
 - trvání slabik
 - reakční čas

Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

- H0: mezi délkou iniciální fráze a přítomností klitika není vztah
- H1: iniciální fráze, po které následuje klitikon, je kratší než fráze, po níž klitikon nenásleduje

Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

- H0: mezi délkou iniciální fráze a přítomností klitika není vztah
- H1: iniciální fráze, po které následuje klitikon, je kratší než fráze, po níž klitikon nenásleduje

- pouze věty obsahující klitikon

Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

- H0: mezi délkou iniciální fráze a přítomností klitika není vztah
- H1: iniciální fráze, po které následuje klitikon, je kratší než fráze, po níž klitikon nenásleduje

- pouze věty obsahující klitikon
- délka fráze měřena v počtu písmen

Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

	prům. délka	sd
Li_P	4.82	2.43
Li_N	9.54	6.23

Příklad – délka iniciální fráze a (ne)přítomnost
klitika po této frázi

- jaký test zvolit?

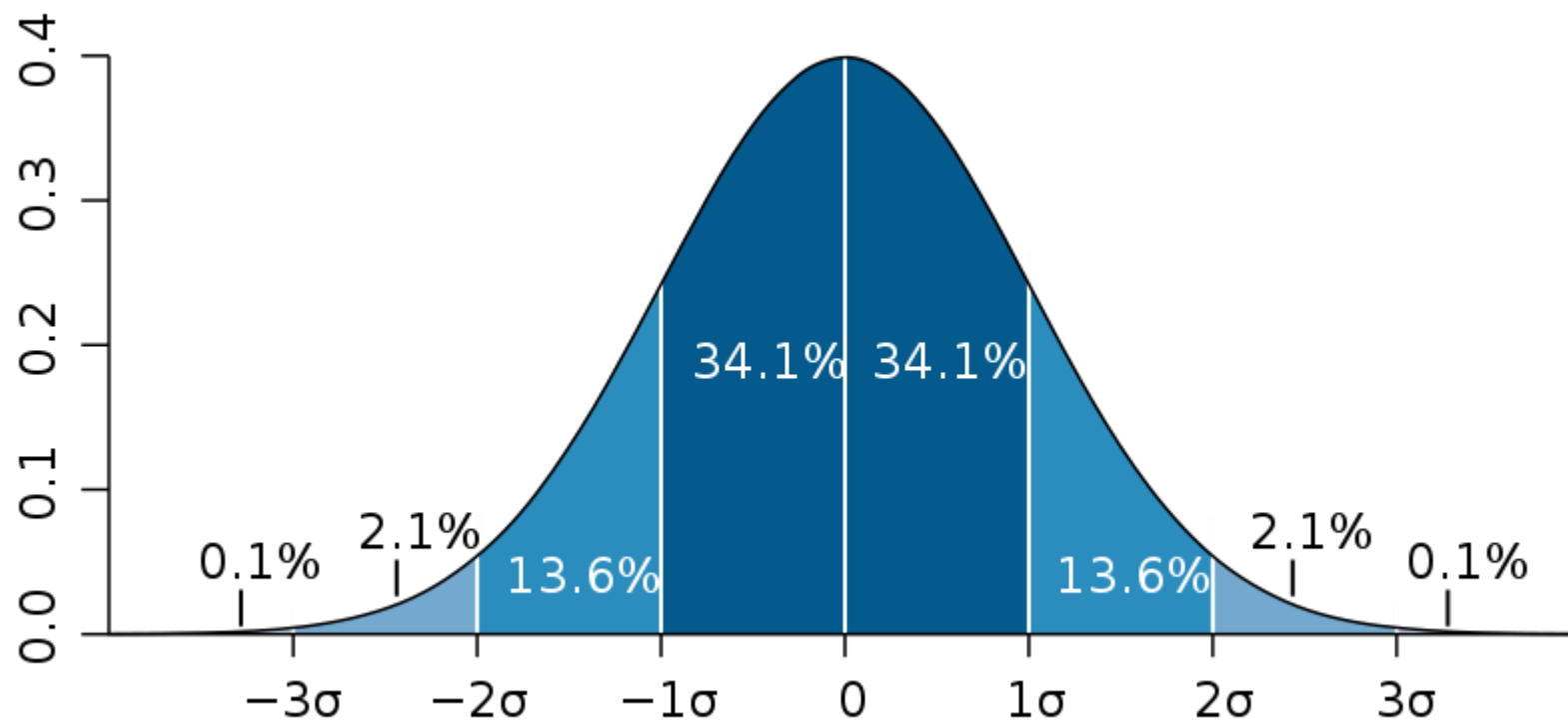
Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

- jaký test zvolit?
- normalita rozdělení dat

Rozdělní (distribuce) dat a jeho interpretace

- statistické testy
 - povaha rozdělní zásadním faktorem pro výběr testu
- parametrické testy
 - předpokládají normální rozdělení
- neparametrické testy
 - nepředpokládají normální rozdělení

Normální rozdělení



Normální rozdělení

- v intervalu od -1σ do 1σ se nachází cca $2/3$ všech hodnot (68,27 %)
- v intervalu od -2σ do 2σ se nachází cca $19/20$ všech hodnot (95,4 %)
- v intervalu od -3σ do 3σ se nachází téměř všechny hodnoty (99,73 %)

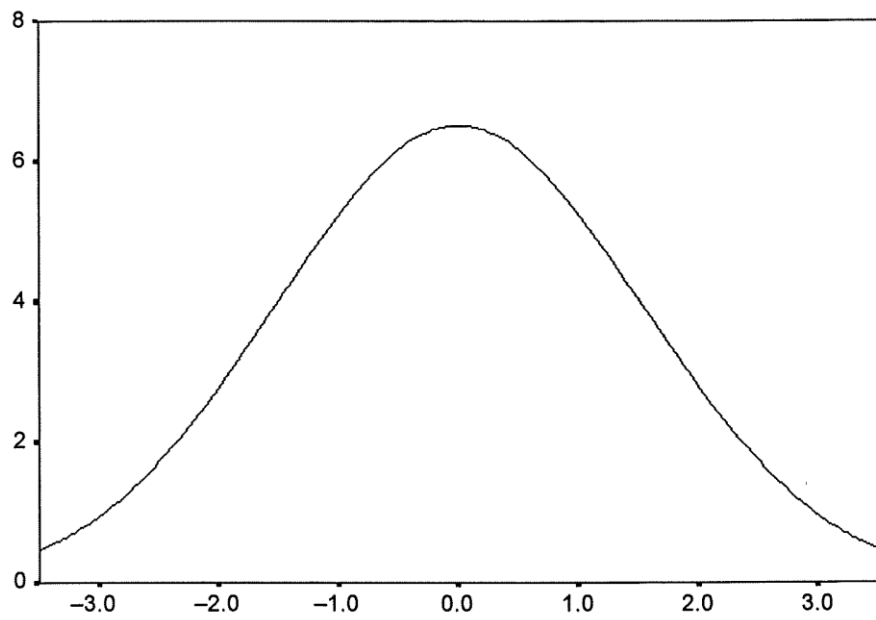


Figure 1.7 A normal curve (mean = 0, SD = 1.53).

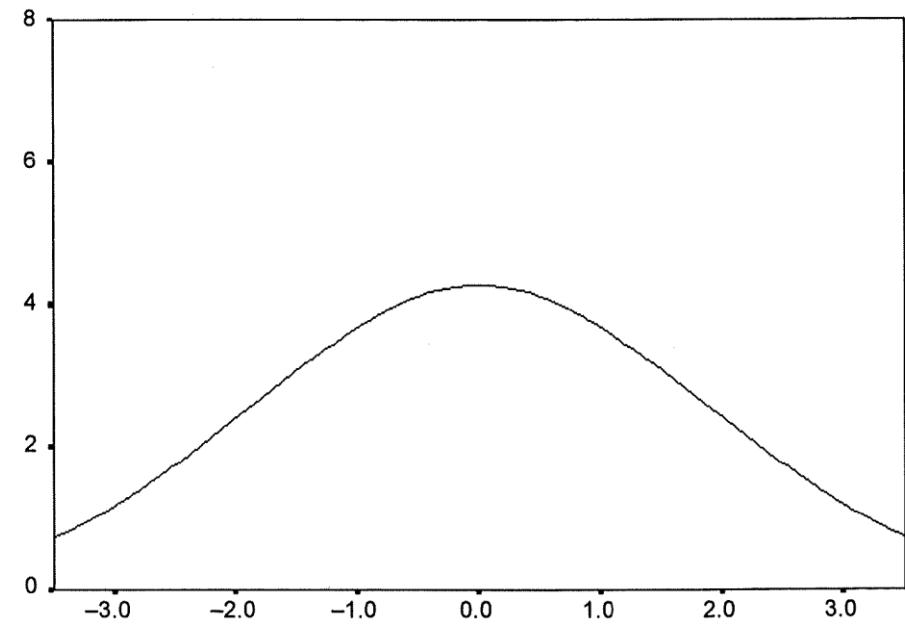


Figure 1.8 A normal curve (mean = 0, SD = 1.86).

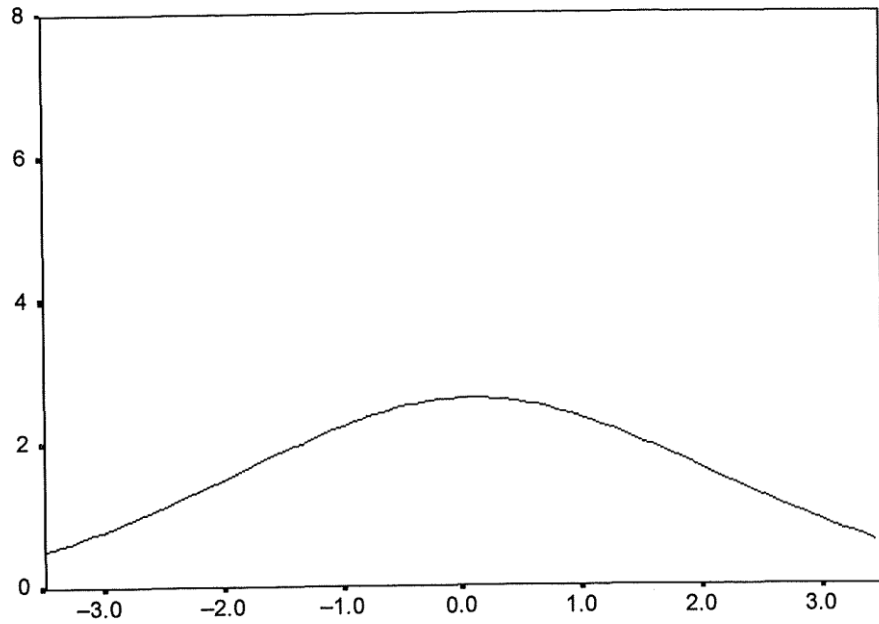


Figure 1.9 A normal curve (mean = 0, SD = 1.98).

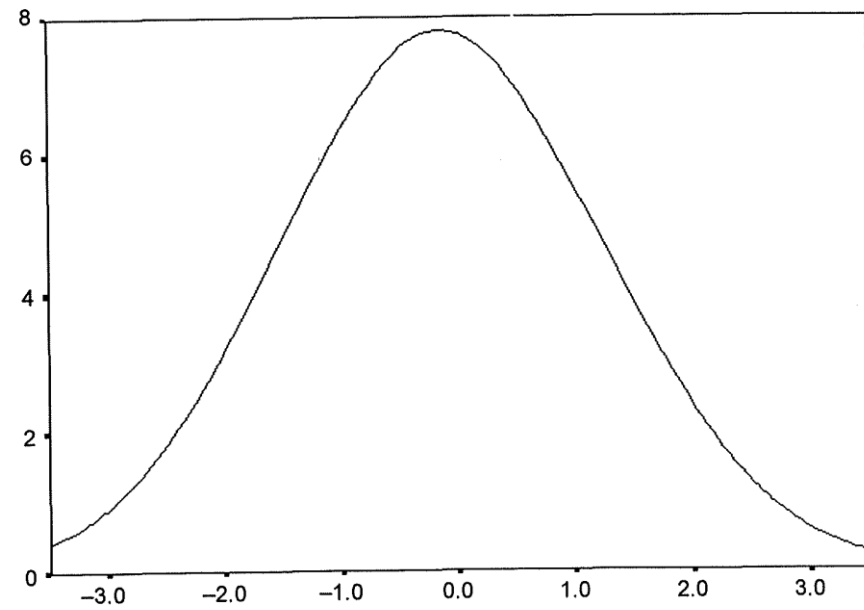


Figure 1.10 A normal curve (mean = 0, SD = 1.38).

Šikmá rozdělení

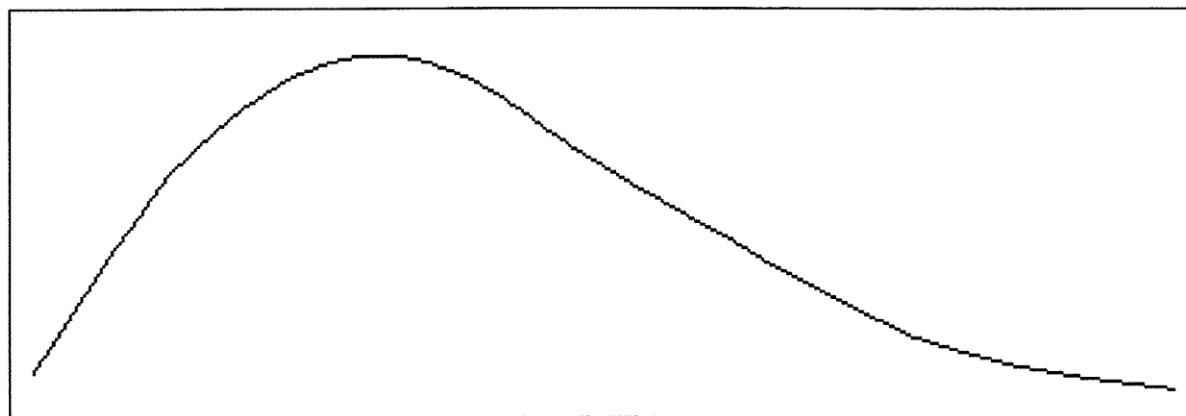


Figure 1.13 A positively skewed distribution.

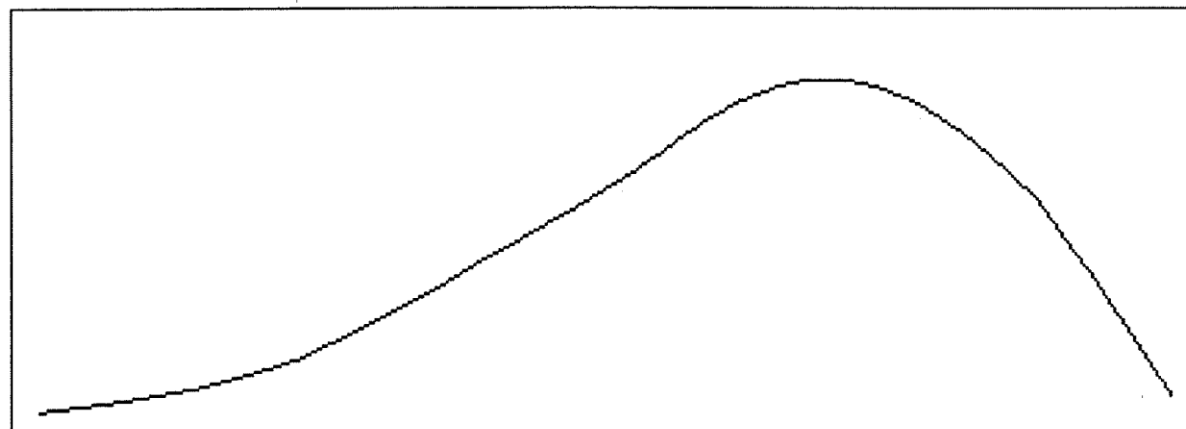


Figure 1.14 A negatively skewed distribution.

Šikmá rozdělení

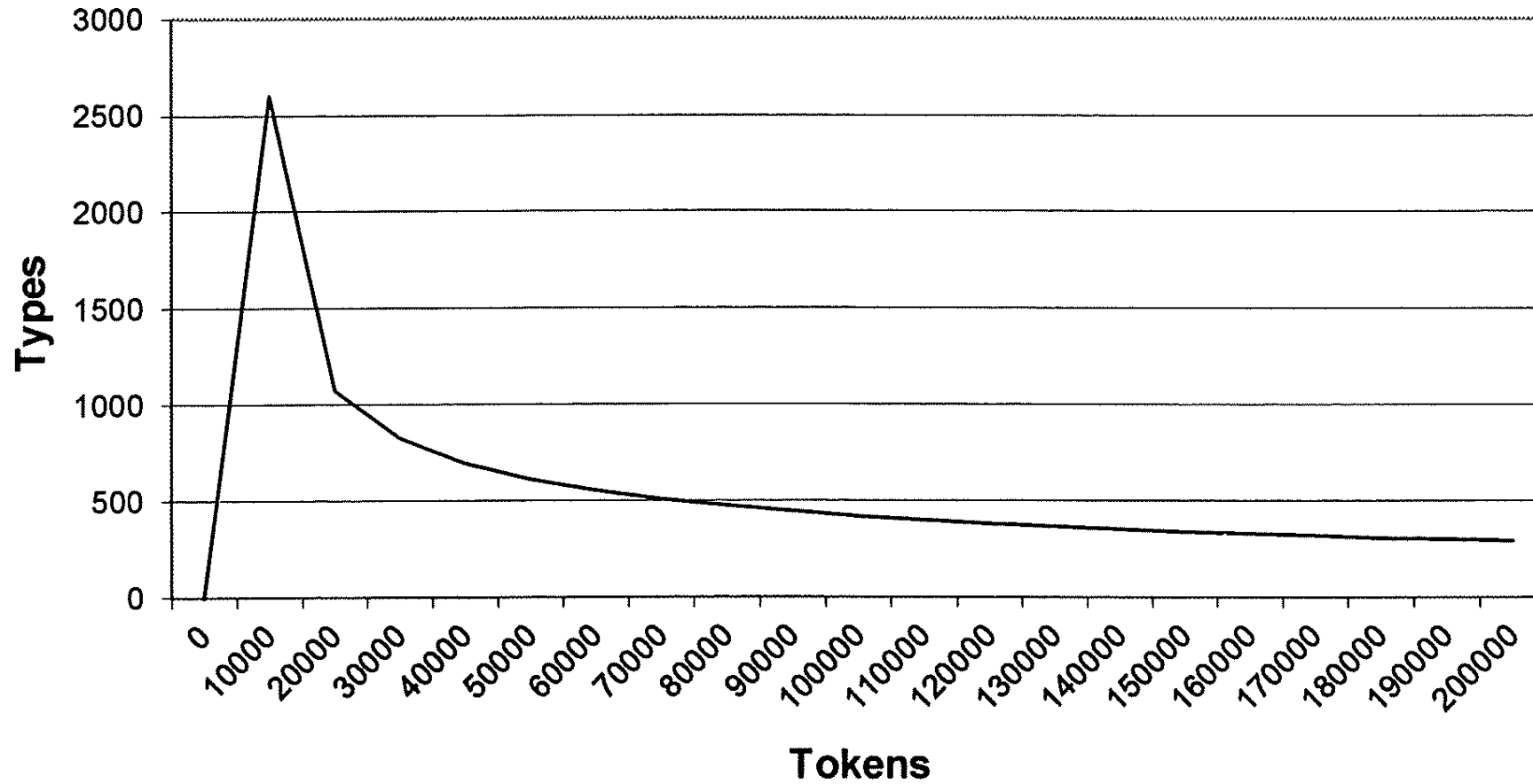
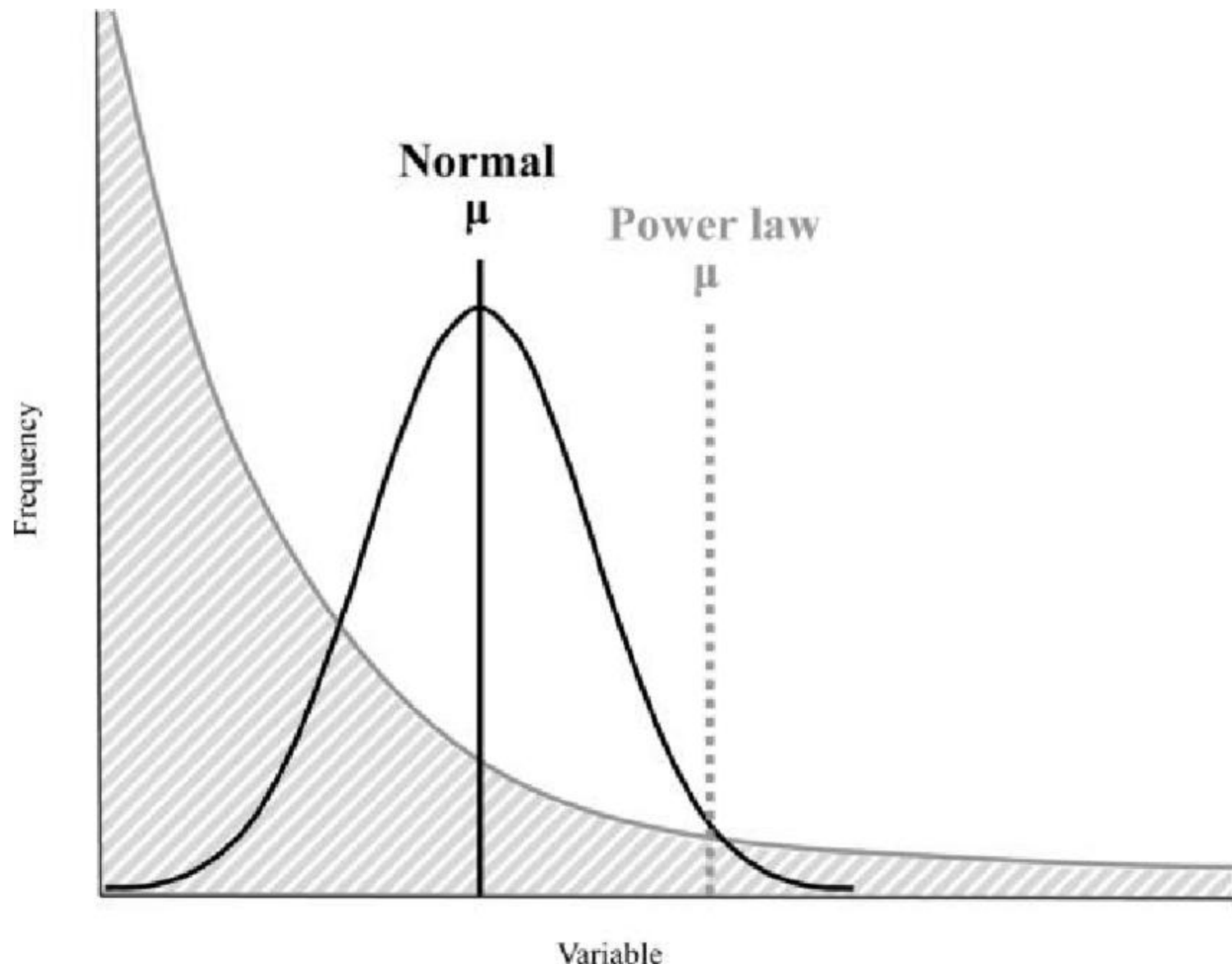
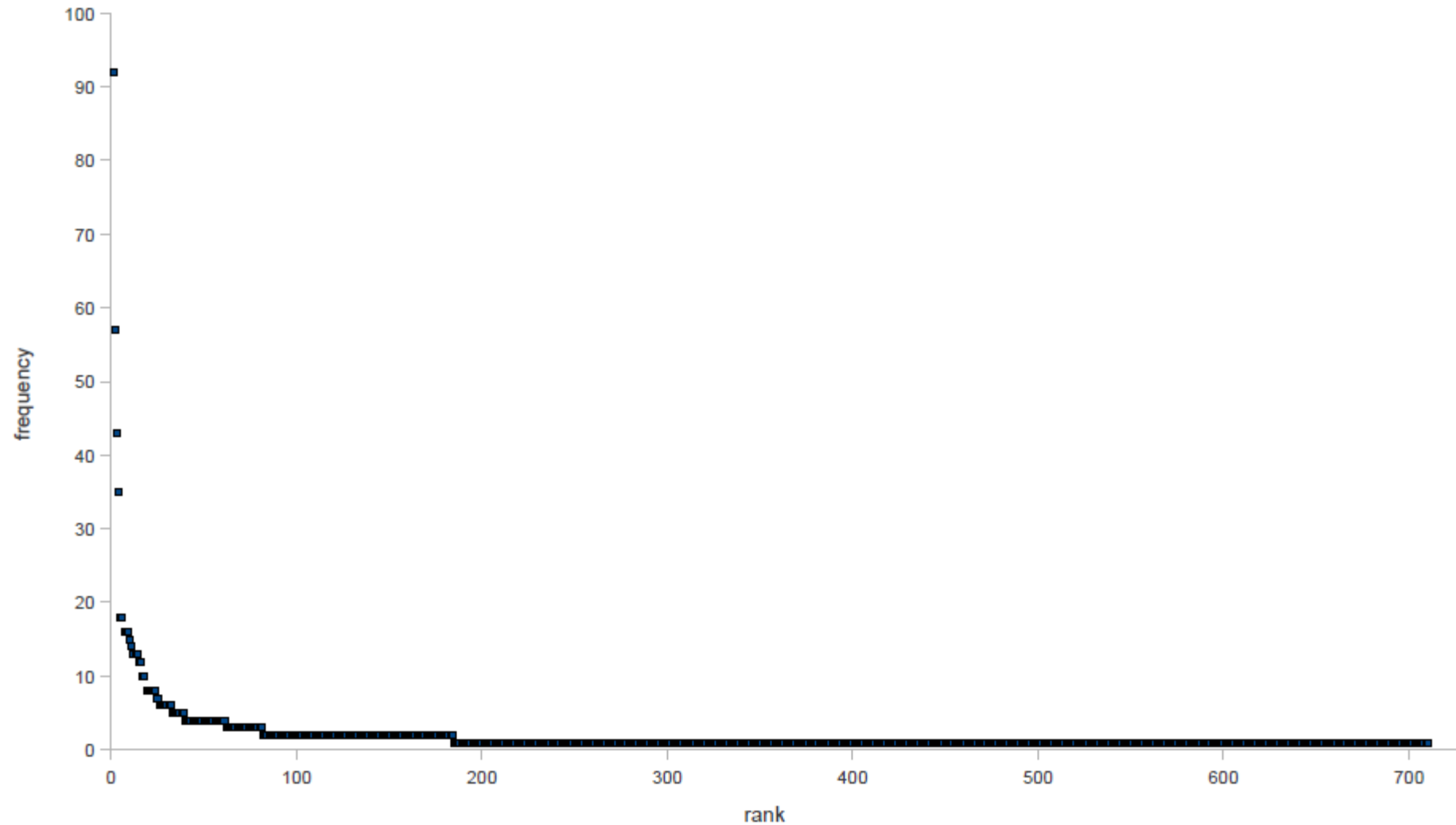
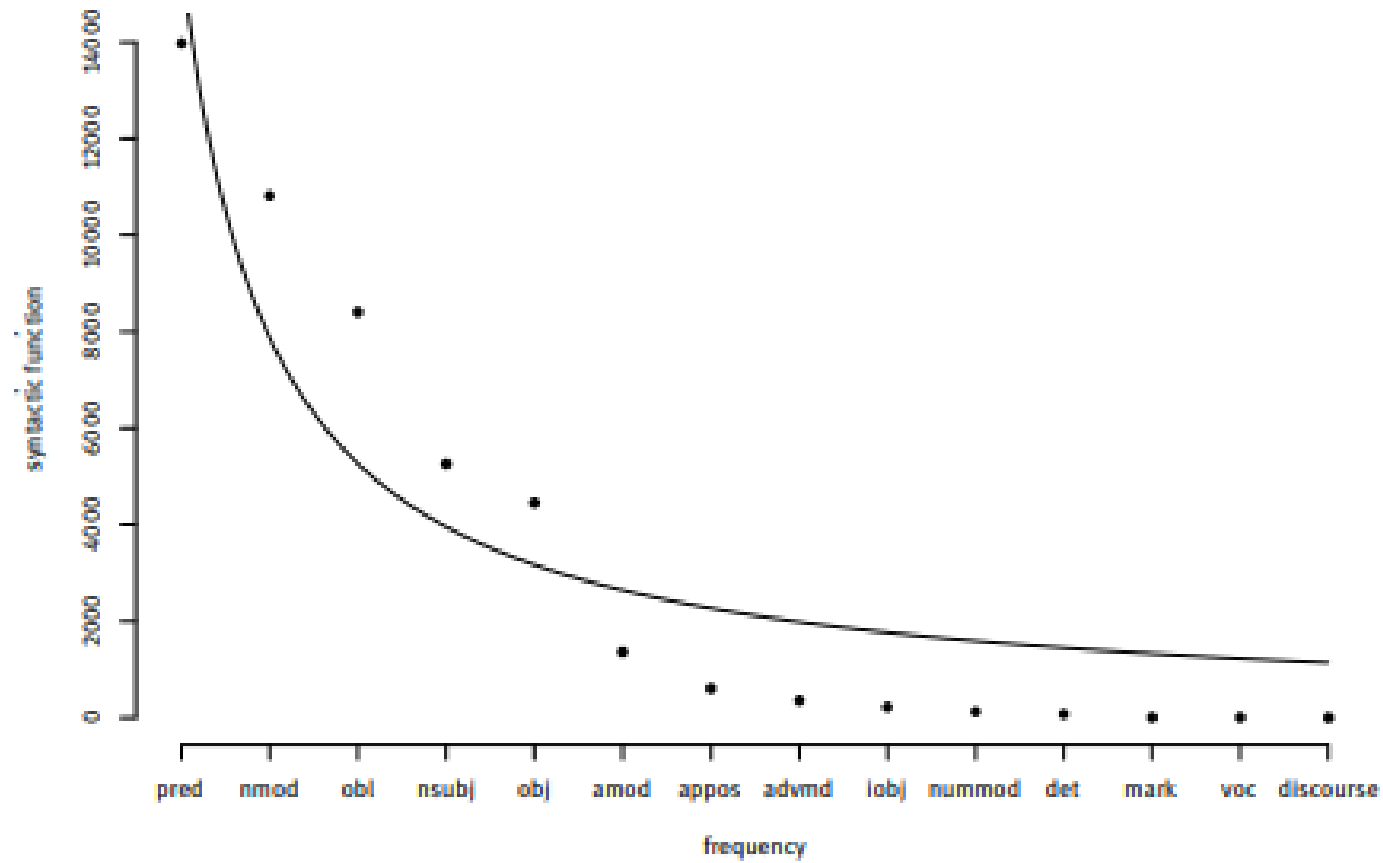


Figure 1.12 Distribution of newly found lexical items throughout a text.



- B. Hrabal: *Zavražděný kohout* ($N = 1435$, $V = 710$, $TTR = V / N = 0.49$)





Čech, R., Milička, J., Mačutek, J., Koščová, M., Lopatková, M. (2018). Quantitative Analysis of Syntactic Dependency in Czech. In Jiang, J., Liu, H. (eds.). Quantitative Analysis of Dependency Structures. De Gruyter, 53-70.

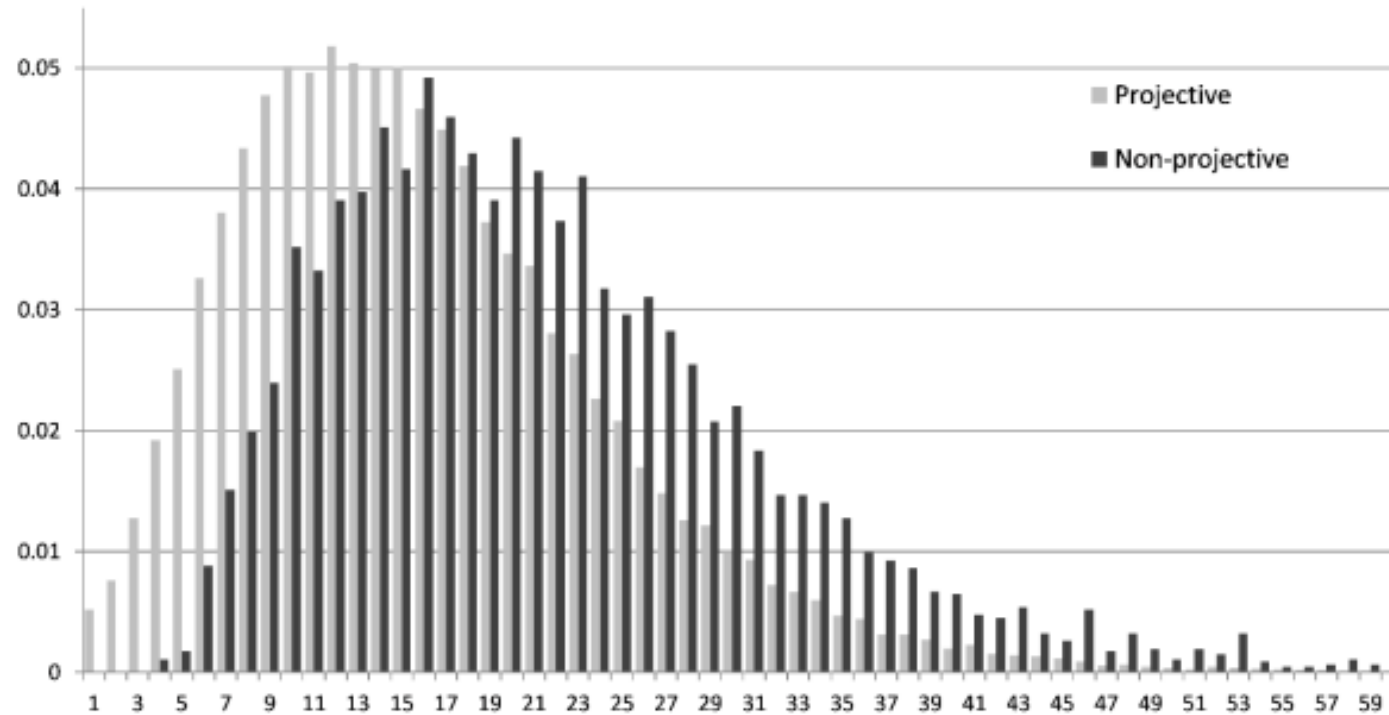


Figure 3. Relative frequencies of lengths of projective (black) and non-projective (grey) sentences in the Czech treebank.

Mačutek, J., Čech, R., Milička, J. (2019). Length of non-projective sentences: A pilot study using a Czech UD treebank. Proceedings of the First Workshop on Quantitative Syntax (Quasy, SyntaxFest 2019), Association for Computational Linguistics, Paris, 110-117.

Předpoklady pro použití testu

- testy normality

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test
 - Kolmogorov-Smirnov Test of Normality

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test
 - Kolmogorov-Smirnov Test of Normality
- výpočet

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test
 - Kolmogorov-Smirnov Test of Normality
- výpočet
 - R

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test
 - Kolmogorov-Smirnov Test of Normality
- výpočet
 - R
 - online kalkulačky

Předpoklady pro použití testu

- testy normality
 - Shapiro-Wilk Normality Test
 - Kolmogorov-Smirnov Test of Normality
- výpočet
 - R
 - online kalkulačky
 - <http://www.statskingdom.com/320ShapiroWilk.html>
 - <https://www.socscistatistics.com/tests/kolmogorov/default.aspx>

Příklad

- hypotetické délky slov v textu

- A: 2, 4, 3, 2, 5, 4, 1, 3, 6, 7

- B: 2, 1, 2, 7, 1, 2, 2, 3, 2, 5

- <http://www.statskingdom.com/320ShapiroWilk.html>

Volba testu

- pokud data normálně rozdělena:
 - t-test
 - R
 - <https://www.socscistatistics.com/tests/studentttest/default2.aspx>
- pokud data neodpovídají normálnímu rozdělení
 - Wilcoxon Signed-Ranks Test (pro spárovaná data)
 - R
 - <https://www.socscistatistics.com/tests/signedranks/default2.aspx>
 - Mann-Whitney U Test Calculator (různé počty hodnot)
 - https://www.statskingdom.com/170median_mann_whitney.html

Příklad

- hypotetické délky slov v textu
 - A: 2, 4, 3, 2, 5, 4, 1, 3, 6, 7
 - B: 2, 1, 2, 7, 1, 2, 2, 3, 2, 5
- normalita
 - <http://www.statskingdom.com/320ShapiroWilk.html>
- test
 - Mann-Whitney U Test Calculator (různé počty hodnot)
 - https://www.statskingdom.com/170median_mann_whitney.html

Příklad – délka iniciální fráze a (ne)přítomnost klitika po této frázi

	prům. délka	sd
Li_P	4.82	2.43
Li_N	9.54	6.23

p-value < 0.001

Příklad – délka frází a (ne)přítomnost klitika

	prům. délka	sd
Ln_N	6.42	2.04
Li_N	9.54	6.23

Příklad – délka frází a (ne)přítomnost klitika

	prům. délka	sd
Ln_N	6.42	2.04
Li_N	9.54	6.23

p-value = 0.28

Cvičení

- porovnejte průměrné délky slov ve dvou textech
 - 5gr_Marketa_Pohoroma_na_silnici.txt
 - 7gr_Anezka_Ohen.txt
-
- hypotéza: žák vyššího ročníku bude používat v průměru delší slova

Cvičení

- porovnejte průměrné délky slov ve dvou textech
- 5gr_Marketa_Pohoroma_na_silnici.txt
- 7gr_Anezka_Ohen.txt

- hypotéza: žák vyššího ročníku bude používat v průměru delší slova

- problémy?

Cvičení

- porovnejte průměrné délky slov ve dvou textech
- 5gr_Marketa_Pohoroma_na_silnici.txt
- 7gr_Anezka_Ohen.txt

- hypotéza: žák vyššího ročníku bude používat v průměru delší slova

- tokeny vs. typy?

Cvičení

- vytvoření seznamu slov
 - <https://ezcalc.me/word-frequency-counter/>
- uložíme do Excelu
- vypočítáme délky slov

	A	B	C	D
1				
2	a	11	=DÉLKA(A2)	
3	na	10		
4	se	8		
5	lukáš	6		
6	ic	5		

Cvičení

- průměr, medián:
 - <https://www.socscistatistics.com/descriptive/averages/default.aspx>
- směrodatná odchylka:
 - <https://www.socscistatistics.com/descriptive/std/default.aspx>
- přehledně zpracovat + např. krabicový graf

Cvičení

- testujte normalitu rozdělení
- <http://www.statskingdom.com/320ShapiroWilk.html>

Cvičení

- nenormální rozdělení → neparametrický test
- Mann-Whitney U Test Calculator (různé počty hodnot)
 - <https://www.socscistatistics.com/tests/mannwhitney/default2.aspx>

Korelace

Korelace

- vzájemný vztah mezi dvěma veličinami

Korelace

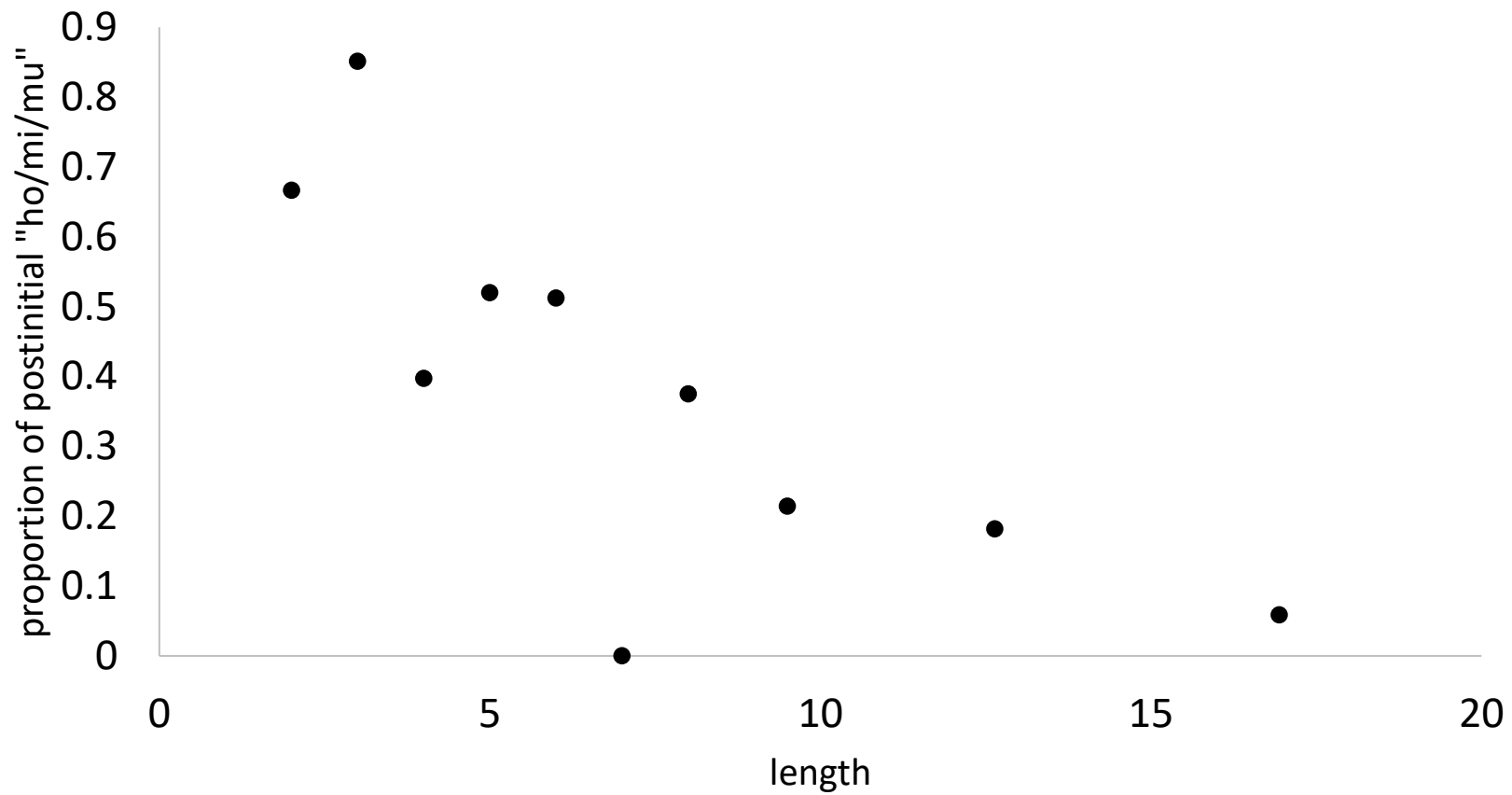
- vzájemný vztah mezi dvěma veličinami
- korelace \neq kauzalita

Korelace

- vzájemný vztah mezi dvěma veličinami
- korelace \neq kauzalita
 - více viz heslo „Korelace neimplikuje kauzalitu“
 - https://cs.wikipedia.org/wiki/Korelace_neimplikuje_kauzalitu
 - srov. příklady zde uvedené

Korelace

- vzájemný vztah mezi dvěma veličinami
- korelace \neq kauzalita
 - více viz heslo „Korelace neimplikuje kauzalitu“
 - https://cs.wikipedia.org/wiki/Korelace_neimplikuje_kauzalitu
- korelační koeficient
 - $\langle -1, +1 \rangle$
 - <https://cs.wikipedia.org/wiki/Korelace>



Proportions of postinitial *mi*, *ho*, *mu* – letters, Pearson's correlation: $r = -0,76$,
p-value = 0,01

Korelace - klasifikace

- 0,00 - 0,19 „velmi slabá“
- 0,20 - 0,39 „slabá“
- 0,40 - 0,59 „střední“
- 0,60 - 0,79 „silná“
- 0,80 - 1,00 „velmi silná“

Korelace – statistická významnost

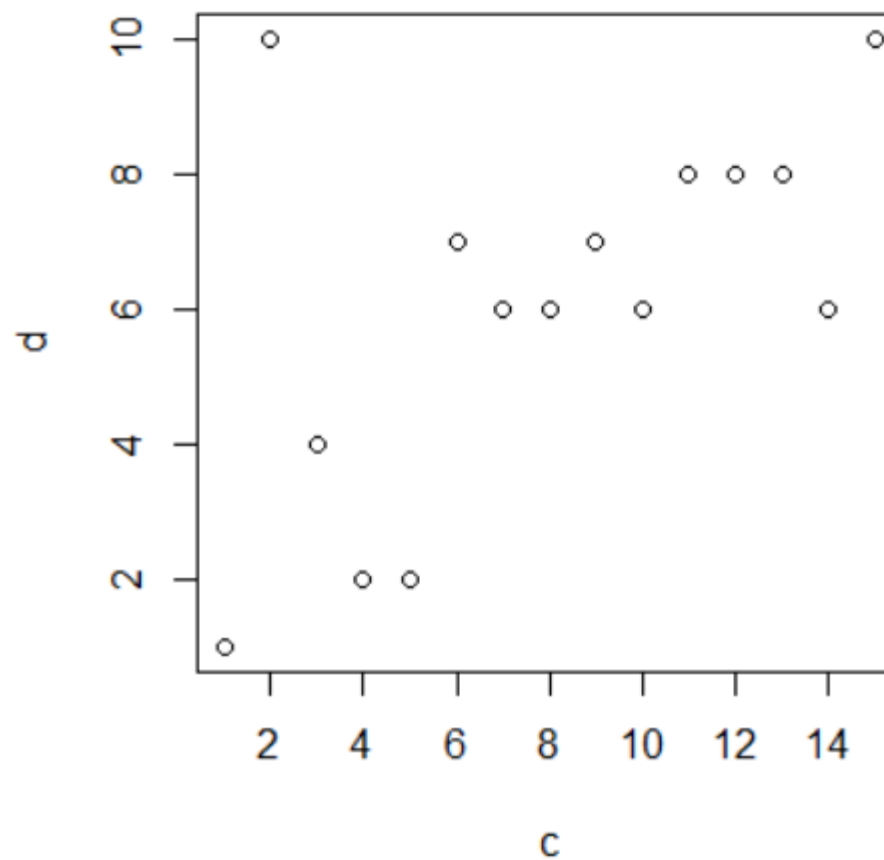
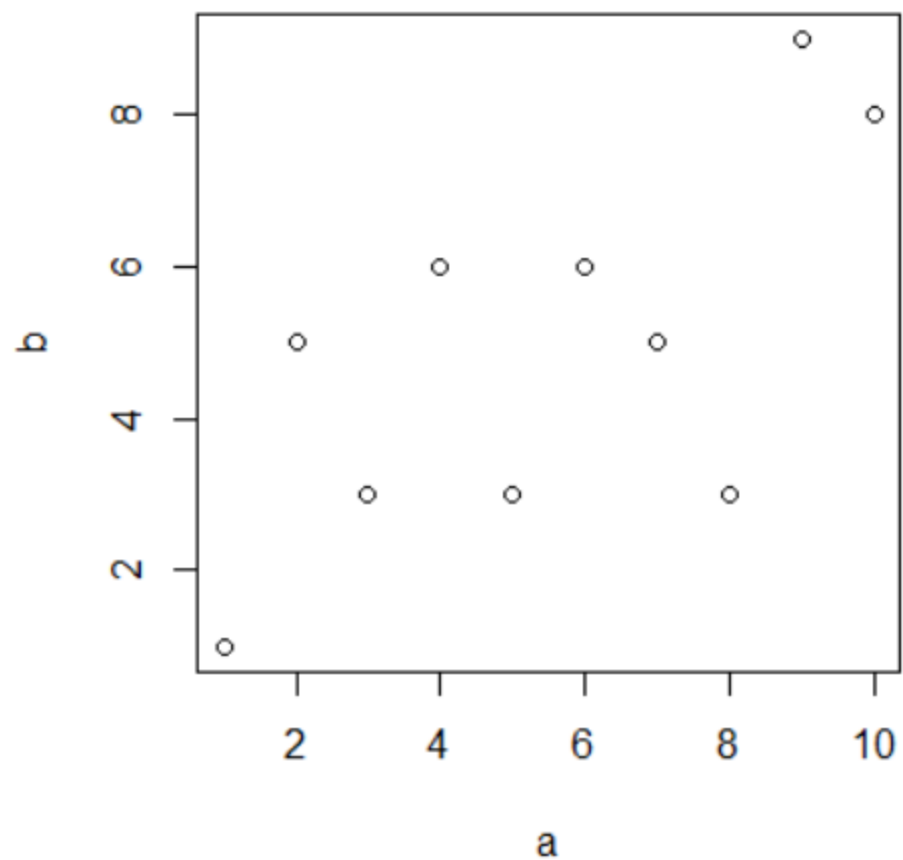
a = (1,2,3,4,5,6,7,8,9,10)

b = (1,5,3,6,3,6,5,3,9,8)

c = (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)

d = (1,10,4,2,2,7,6,6,7,6,8,8,8,6,10)

Korelace – statistická významnost



Korelace – statistická významnost

a = (1,2,3,4,5,6,7,8,9,10)

b = (1,5,3,6,3,6,5,3,9,8)

Kendallův koeficient

tau = 0,471

c = (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)

d = (1,10,4,2,2,7,6,6,7,6,8,8,8,6,10)

Kendallův koeficient

tau = 0,475

Korelace – statistická významnost

a = (1,2,3,4,5,6,7,8,9,10)

b = (1,5,3,6,3,6,5,3,9,8)

Kendallův koeficient

tau = 0,471

p-value = 0,067

c = (1,2,3,4,5,6,7,8,9,10,11,12,13,14,15)

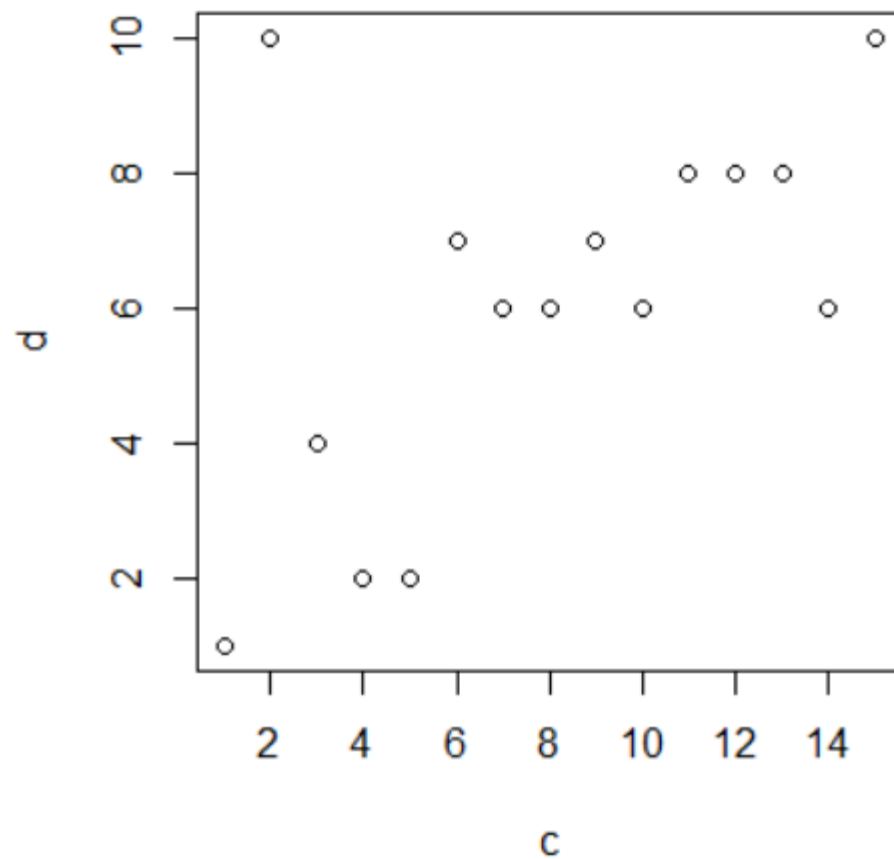
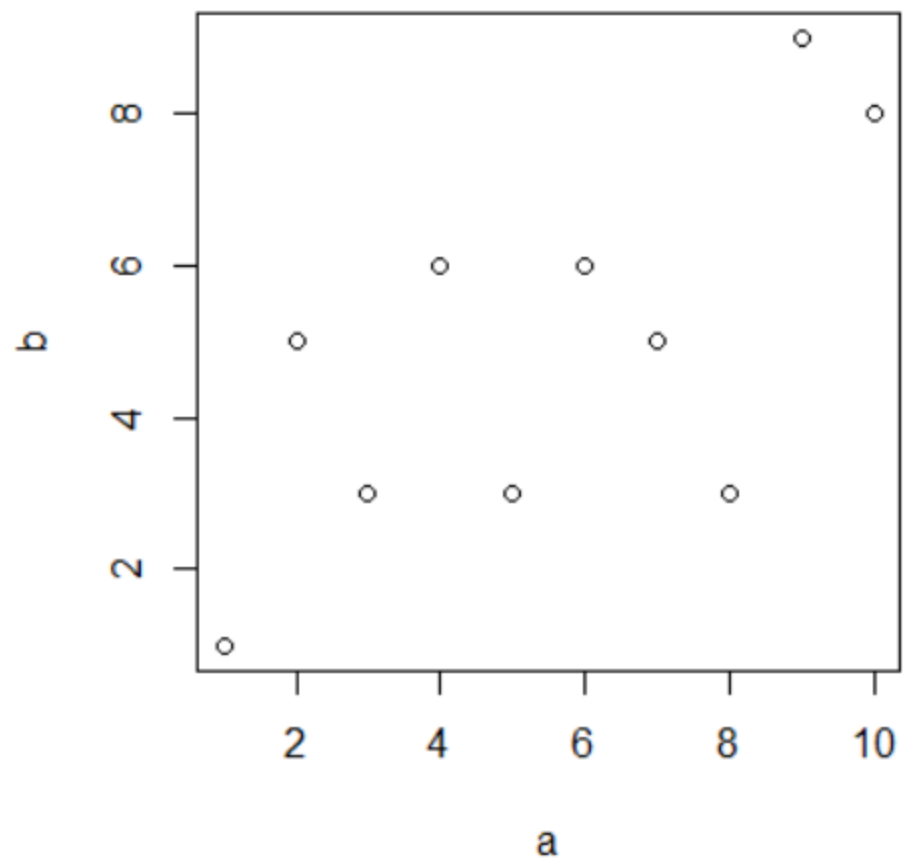
d = (1,10,4,2,2,7,6,6,7,6,8,8,8,6,10)

Kendallův koeficient

tau = 0,475

p-value = 0,017

Korelace – statistická významnost



Korelace – výběr testu

- pokud data normálně rozdělena:
 - Pearsonův korelační koeficient
 - R
 - <https://www.socscistatistics.com/tests/pearson/default.aspx>

Korelace – výběr testu

- pokud data normálně rozdělena:
 - Pearsonův korelační koeficient
 - R
 - <https://www.socscistatistics.com/tests/pearson/default.aspx>
- pokud data neodpovídají normálnímu rozdělení
 - Spearmanův korelační koeficient
 - R
 - [Spearman's Rho \(Correlation\) Calculator](#)
 - Kendallův test
 - R
 - https://www.wessa.net/rwasp_kendall.wasp

Úkol 1

- porovnejte průměrné délky slov ve dvou textech
 - prvním textem bude báseň
 - druhým textem bude část odborného textu (např. první odstavec, v případě, že bude krátký, tak první dva odstavce)
- zjistěte
 - zda délky odpovídají/neodpovídají normálnímu rozdělení
 - zda je rozdíl délek statisticky významný, nebo ne
- výsledky přehledně zpracujte
 - popište jazykový materiál
 - formou tabulky uveďte všechny zjištěné hodnoty délek
 - popište postup analýzy (volba testu, měření normality atp.)
 - prezentujte výsledky

Úkol 2

- z předchozího úkolu zjistěte frekvence jednotlivých délek slov (u každého textu zvlášť)
- zkoumejte vztah mezi frekvencí jednotlivých délek a délkami
 - platí vztah, že jsou frekventovaná slova kratší?
 - vytvořte graf, kde budou na ose x jednotlivé délky slov a na ose y jejich frekvence
 - pomocí měření korelace zjistěte, zda je mezi oběma veličinami vztah
- výsledky přehledně zpracujte
 - formou tabulky uveďte všechny zjištěné hodnoty délek
 - popište postup analýzy (volba testu, měření normality atp.)
 - prezentujte výsledky