Kritická práce s daty 3

Radek Čech

Střední hodnoty

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20}

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20} průměr = 5,67

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20} průměr = 5,67

{5,5,6,6,6,6}

Aritmetický průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

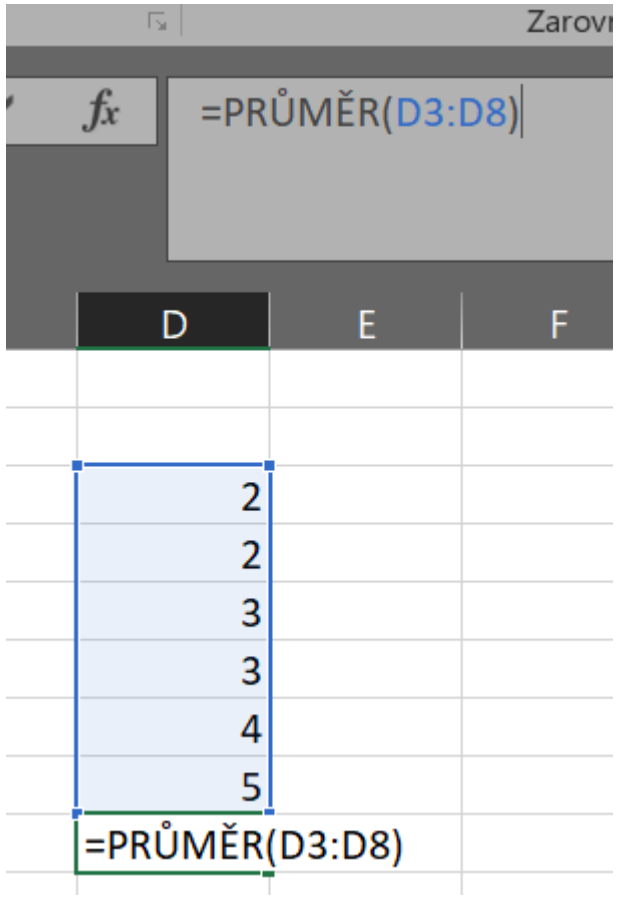
! citlivý na od odlehlé hodnoty

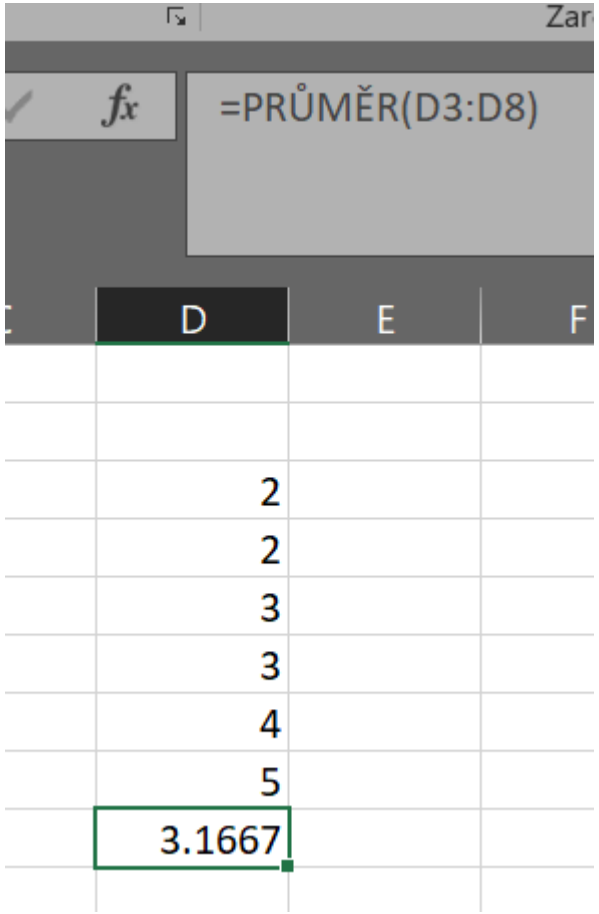
{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20} průměr = 5,67

{5,5,6,6,6,6} průměr = 5,67

Aritmetický průměr v Excelu





Variabilita dat – rozptyl & směrodatná odchylka

- rozptyl
  - střední hodnota kvadrátů odchylek od střední hodnoty

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N - 1}$$

Variabilita dat – rozptyl & směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17

$$\sigma^2 = \frac{(2-3{,}17)^2 + (2-3{,}17)^2 + (3-3{,}17)^2 + (3-3{,}17)^2}{6-1} +$$

$$+ \frac{(4-3{,}17)^2 + (5-3{,}17)^2}{} =$$

$$= \frac{1{,}3689 + 1{,}3689 + 0{,}0289 + 0{,}0289 + 0{,}6889 + 3{,}3489}{5} = \frac{6{,}8334}{5} =$$

$$= 1{,}367$$

Variabilita dat – směrodatná odchylka

směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = 1{,}169$$

Variabilita dat – směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17 SD =1,17

{2,2,3,3,4,20} průměr = 5,67 SD = 7,06

{5,5,6,6,6,6} průměr = 5,67 SD = 0,52

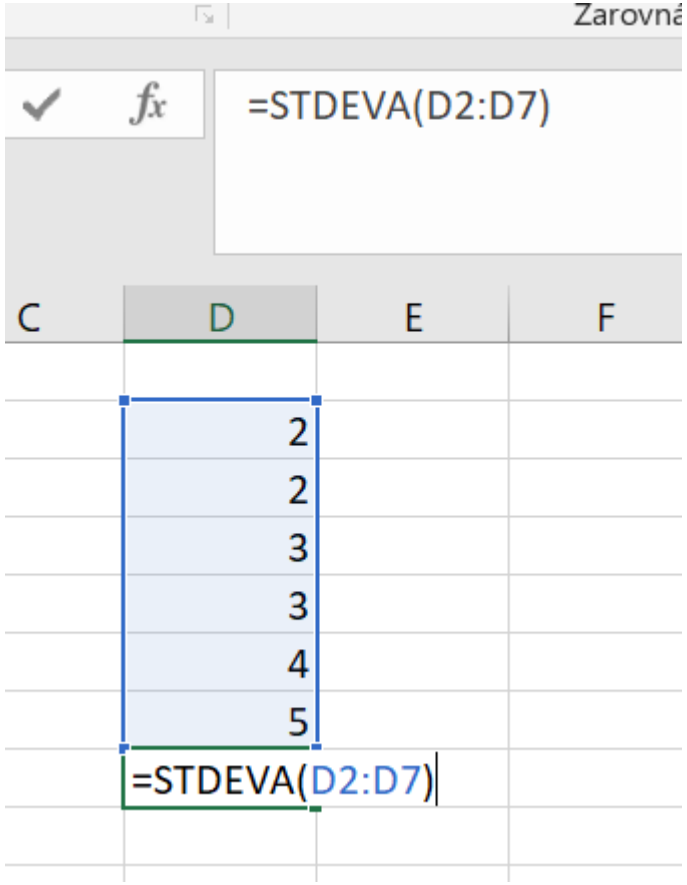
Variabilita dat – směrodatná odchylka

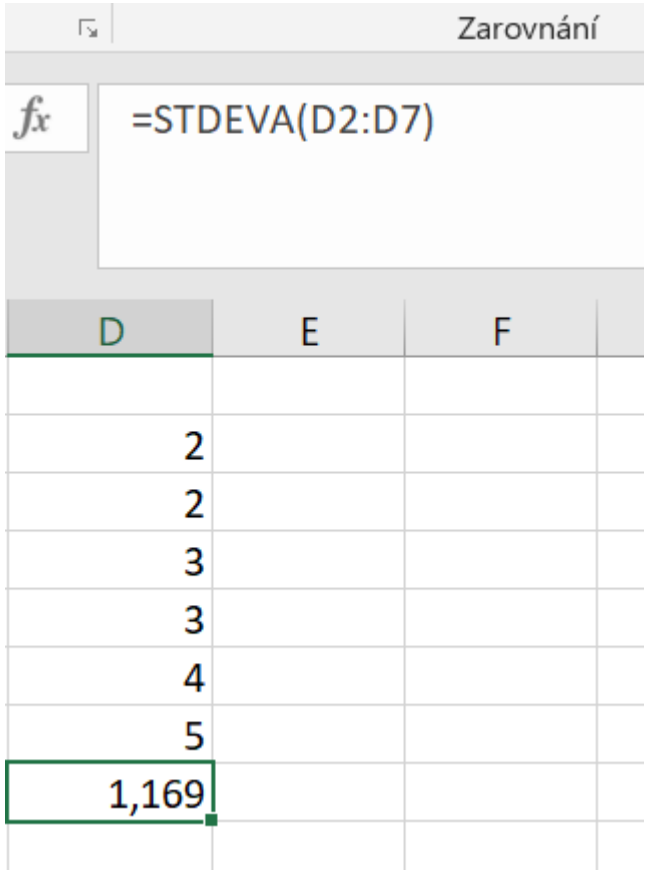
{2,2,3,3,4,5} průměr = 3,17 SD = 1,17

{2,2,3,3,4,20} průměr = 5,67 SD = 7,06

{5,5,6,6,6,6} průměr = 5,67 SD = 0,52

SD v Excelu





více viz https://support.office.com/cs-cz/article/stdeva-funkce-5ff38888-7ea5-48de-9a6d-11ed73b29e9d

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

Slovo vs. mluvní takt

- v čem je rozdíl?

Slovo vs. mluvní takt

- stůl
- na stole
- v domě
- napil se
- napil jsem se
- podali jsme jim ho
- řekl, že přijde

Slovo vs. mluvní takt

- stůl
- na stole
- v domě
- napil se
- napil jsem se
- podali jsme jim ho
- řekl, že přijde

- stůl
- nastole
- vdomě
- napilse
- napiljsemse
- podalijsmejimho
- řekl žepřijde

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

- jazykový materiál
  - Ukradený kaktus (K. Čapek)
  - Žánrové a stylové proměny veřejné jazykové komunikace (J. Kraus)

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

- jazykový materiál
  - Ukradený kaktus (K. Čapek)
  - Žánrové a stylové proměny veřejné jazykové komunikace (J. Kraus)
- segmentace
  - S jako grafická jednotka, délka (L) měřena v počtu slabik
  - MT vymezen podle Palkové (2004), délka (L) měřena v počtu slabik

Porovnání délek – a jeho interpretace

- očekávání
  - S budou kratší než MT
  - délka S a MT bude delší v odborných textech než v beletrii

Porovnání délek – a jeho interpretace

- očekávání
  - S budou kratší než MT
  - délka S a MT bude delší v odborných textech než v beletrii

- jak měřit?

Výsledky – průměrné délky

	L_S	L_{MT}
bel	2	2,89
odb	2,83	3,51

Výsledky – průměrné délky a SD

	L_S	SD_S	L_{MT}	SD_{MT}
bel	2	1,05	2,89	1
odb	2,83	1,4	3,51	1,23

Medián

- rozděluje soubor na dvě stejné poloviny
- není ovlivněn extrémními hodnotami

Medián

- rozděluje soubor na dvě stejné poloviny
- není ovlivněn extrémními hodnotami

{2,2,3,3,4,5} průměr = 3,17 medián = 3

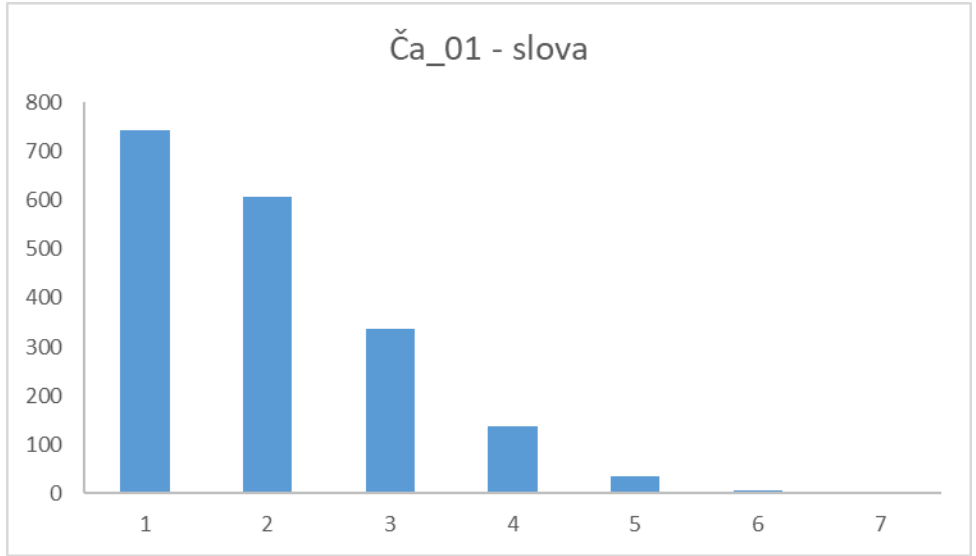
{2,2,3,3,4,20} průměr = 5,67 medián = 3

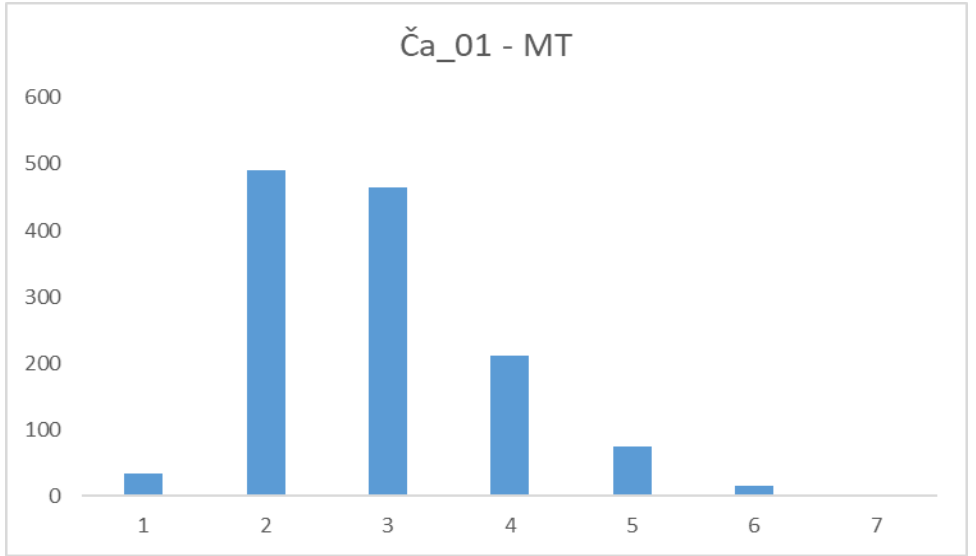
{5,5,6,6,6,6} průměr = 5,67 medián = 6

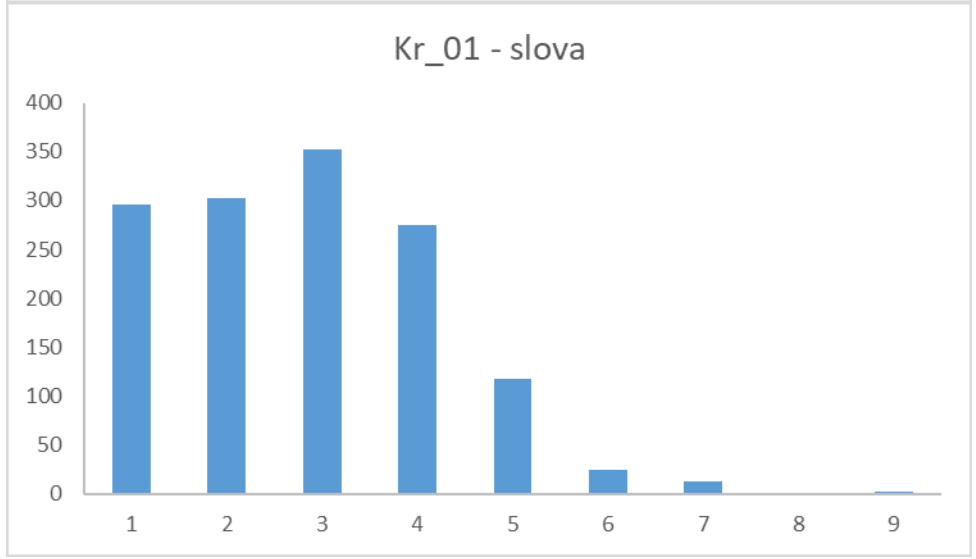
Výsledky – průměrné délky, SD, medián

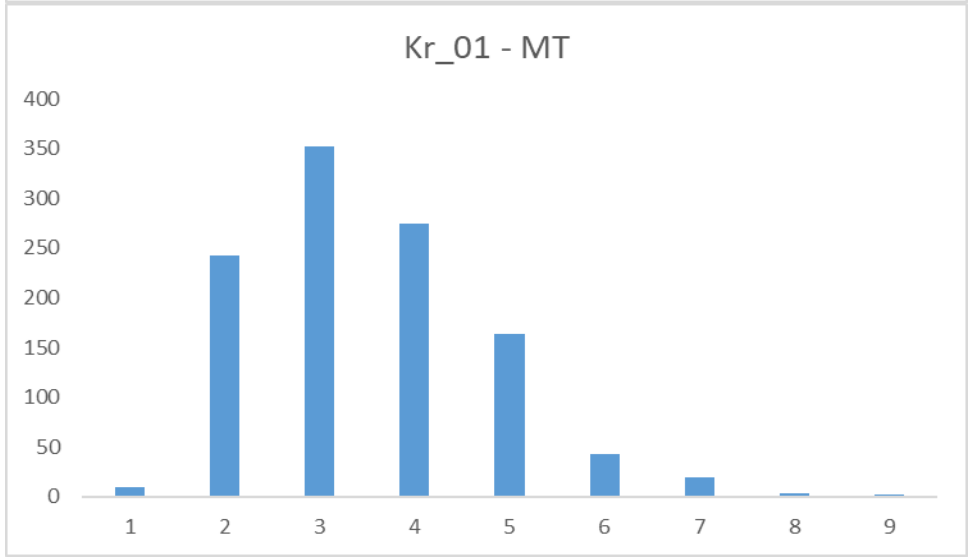
	L_S	SD_S	M_S	L_{MT}	SD_{MT}	M_{MT}
bel	2	1,05	2	2,89	1	3
odb	2,83	1,4	3	3,51	1,23	3

Porovnání délek – jeho interpretace & grafické znázornění





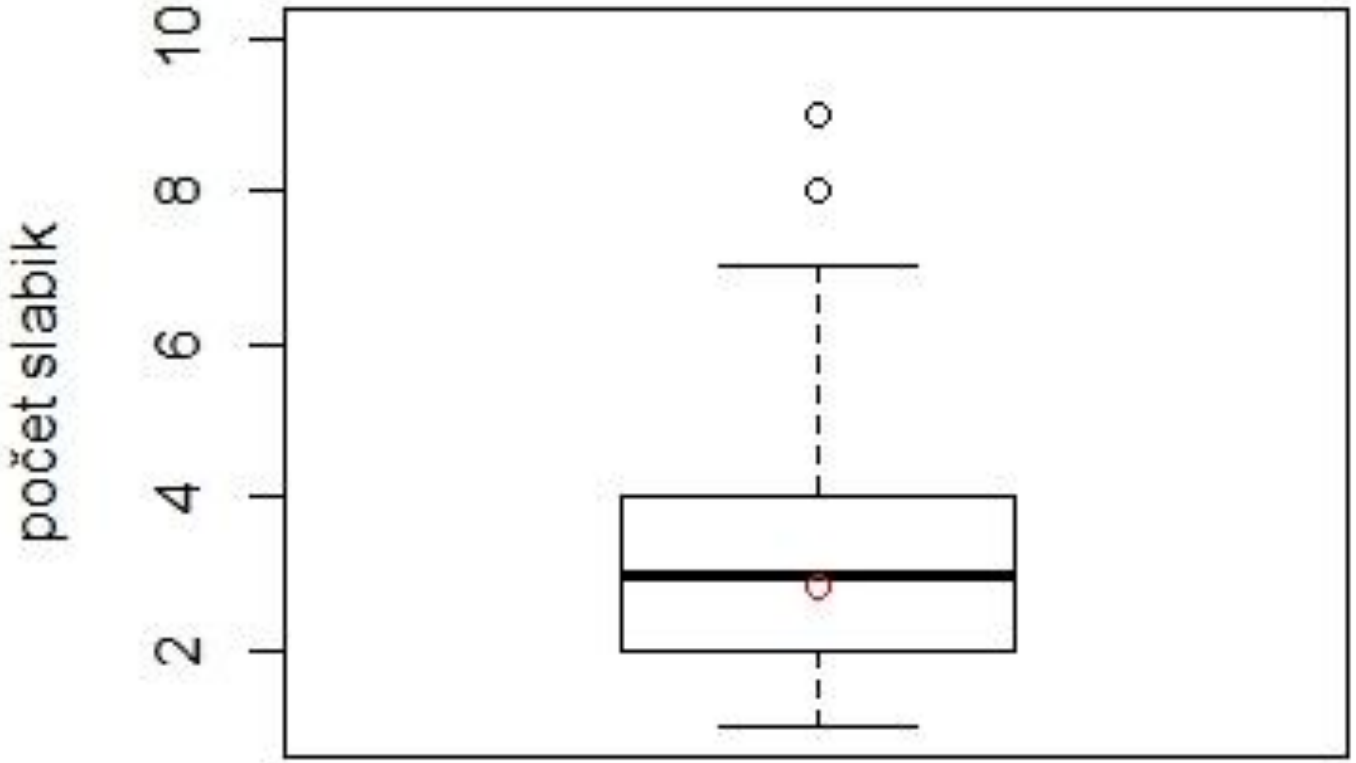




Porovnání délek – jeho interpretace & grafické znázornění

- Kr_01 – MT

- dokážete interpretovat tento graf?

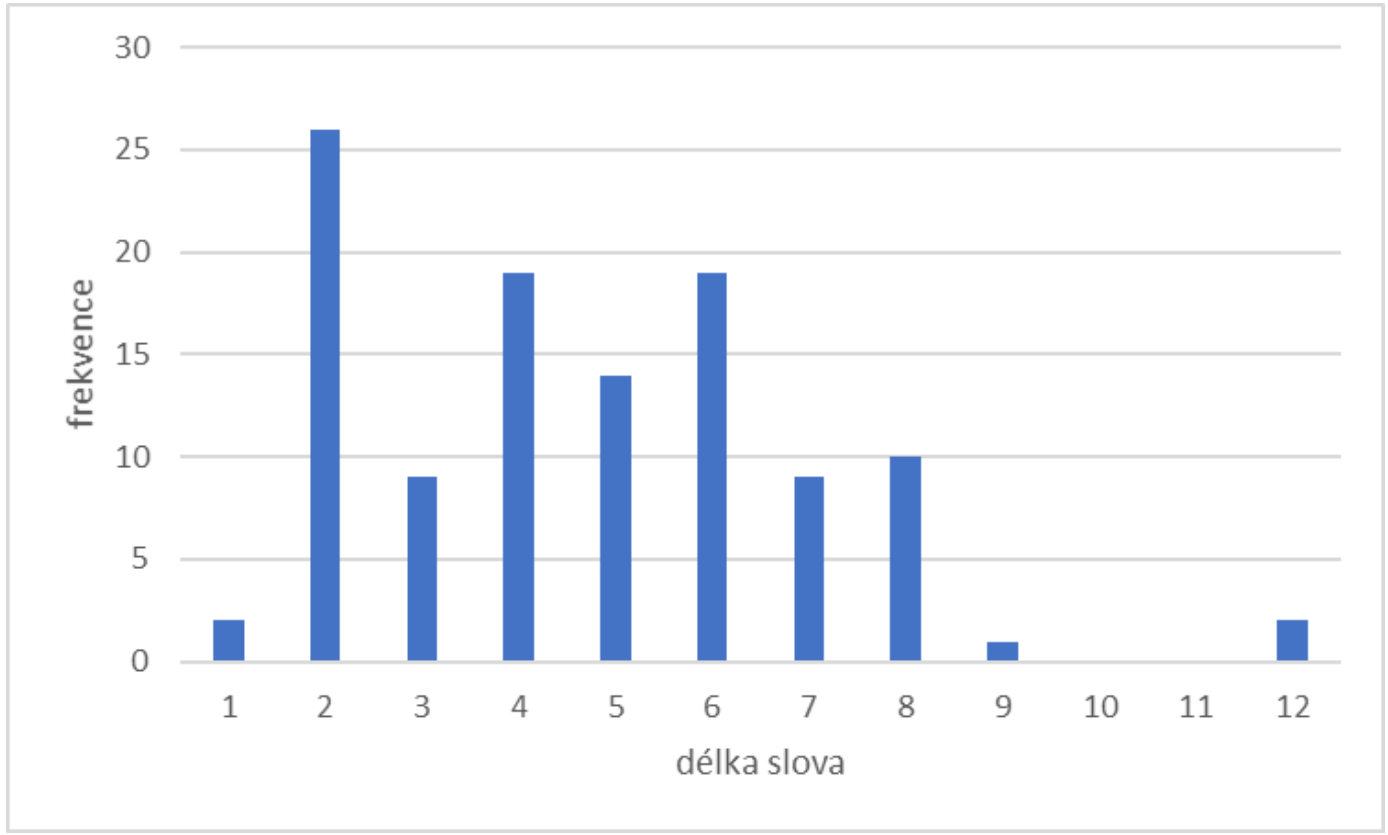


Krabicový graf

- rozložení hodnot

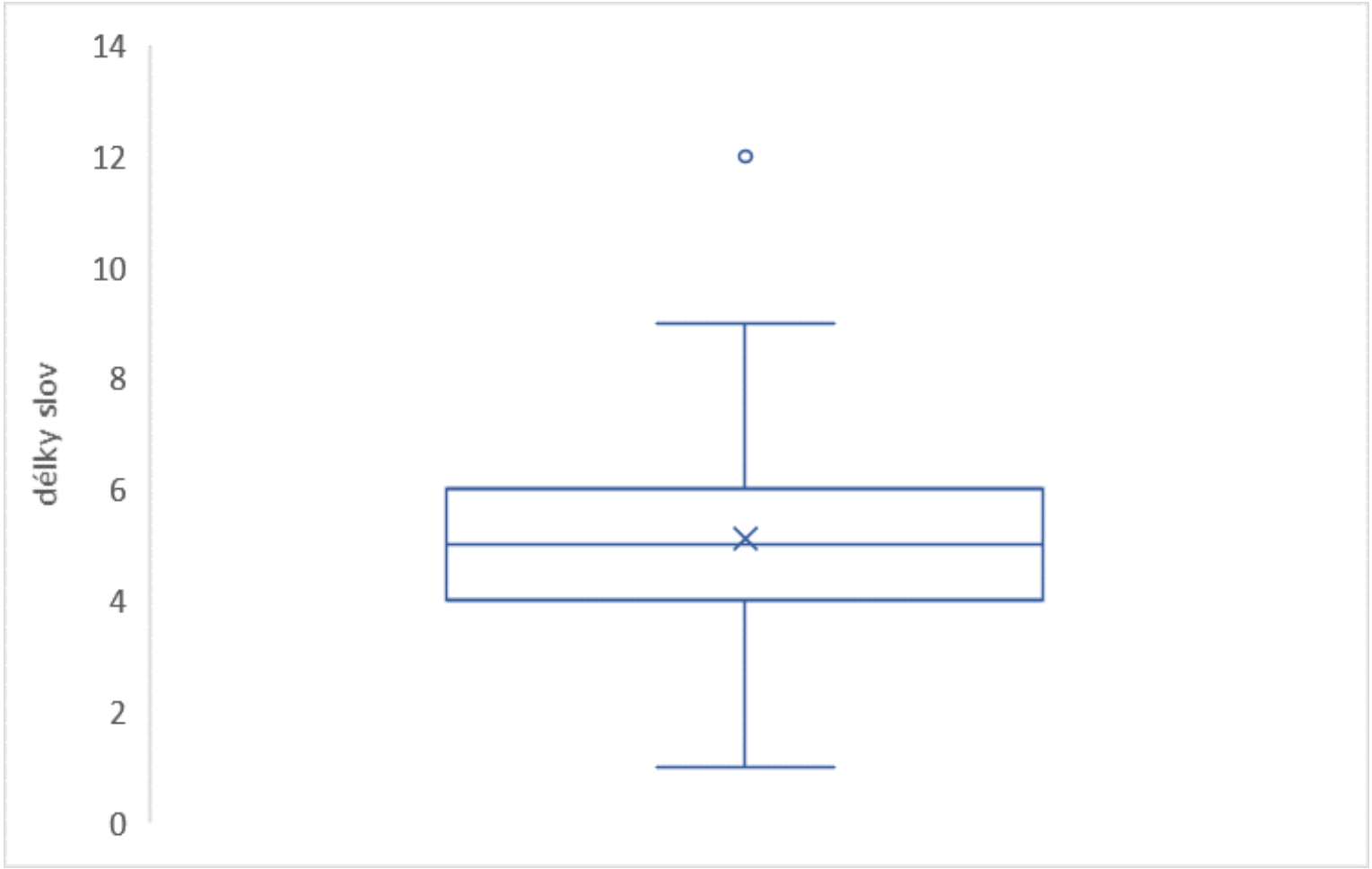
Krabicový graf

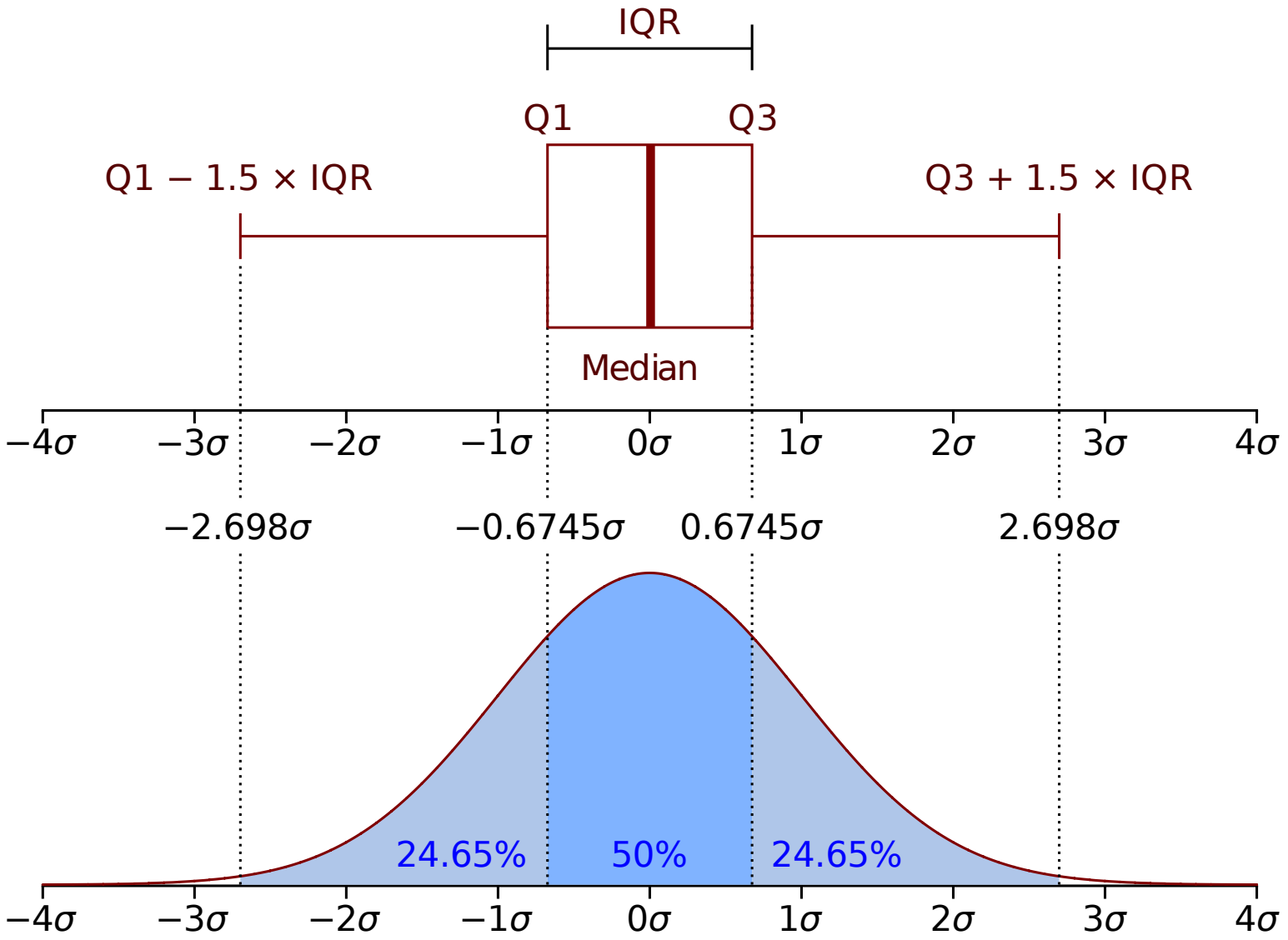
- rozložení hodnot



Krabicový graf

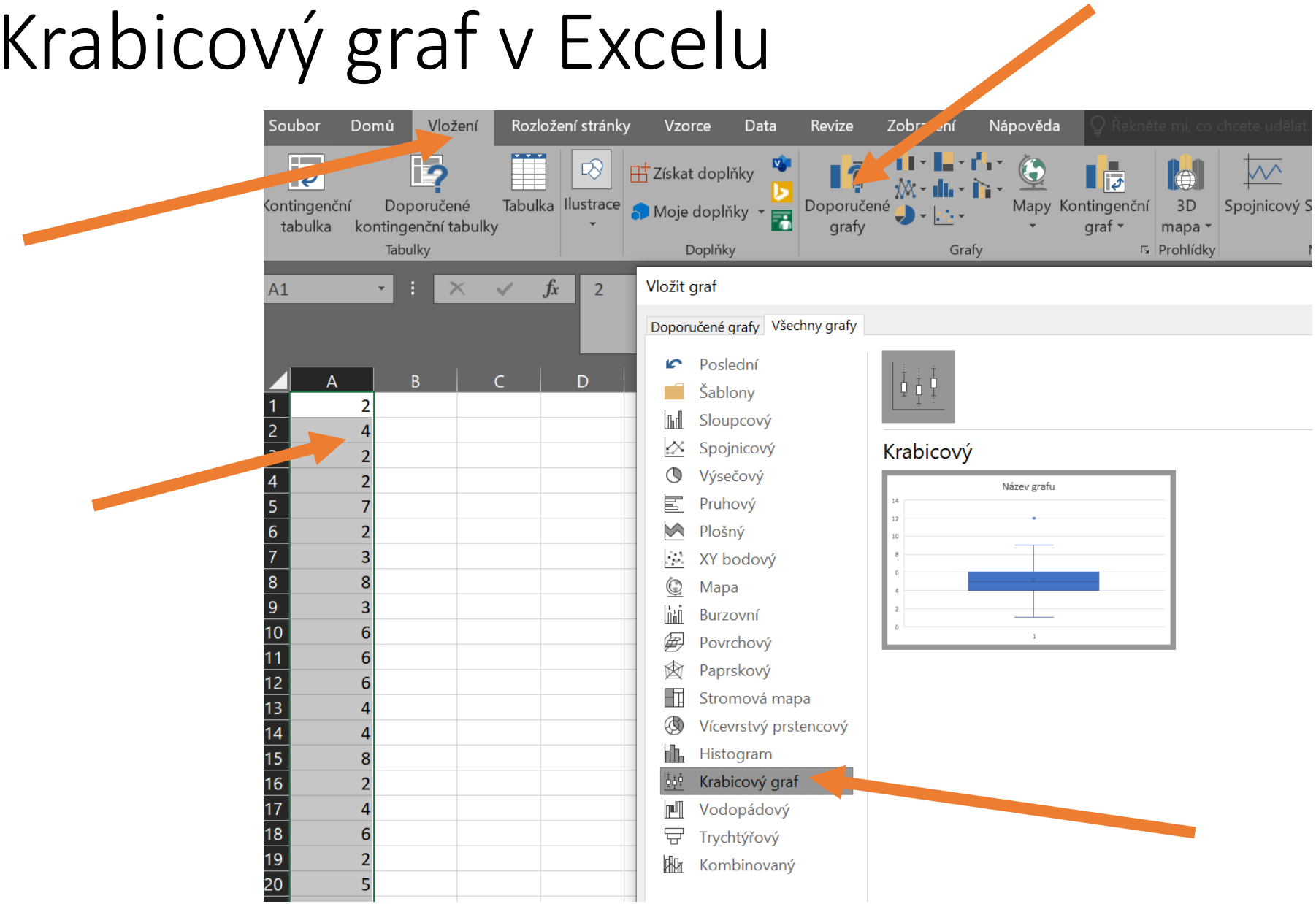
- rozložení hodnot



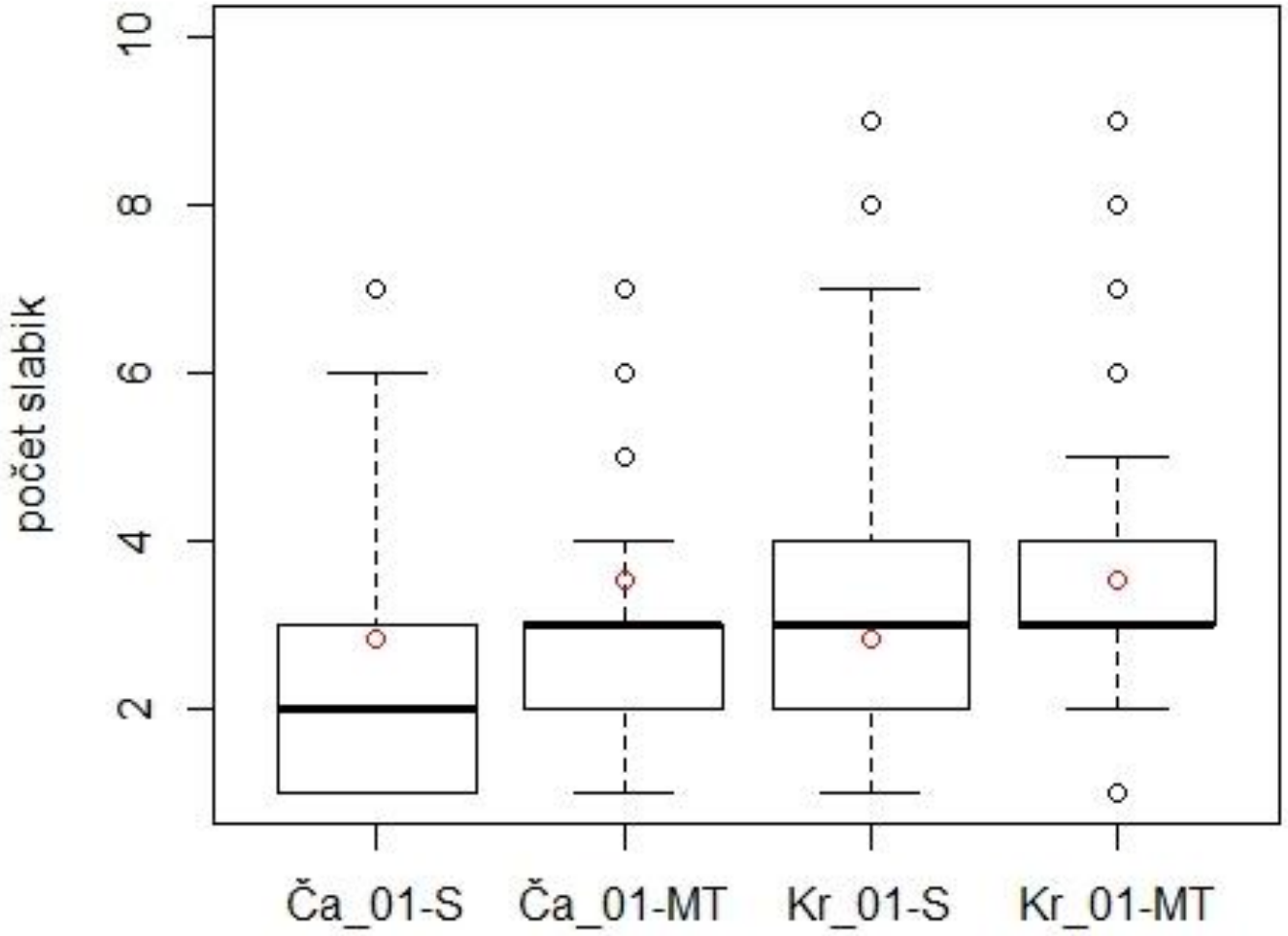


- Autor: Jhguch at en.wikipedia, CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=14524285

Krabicový graf v Excelu



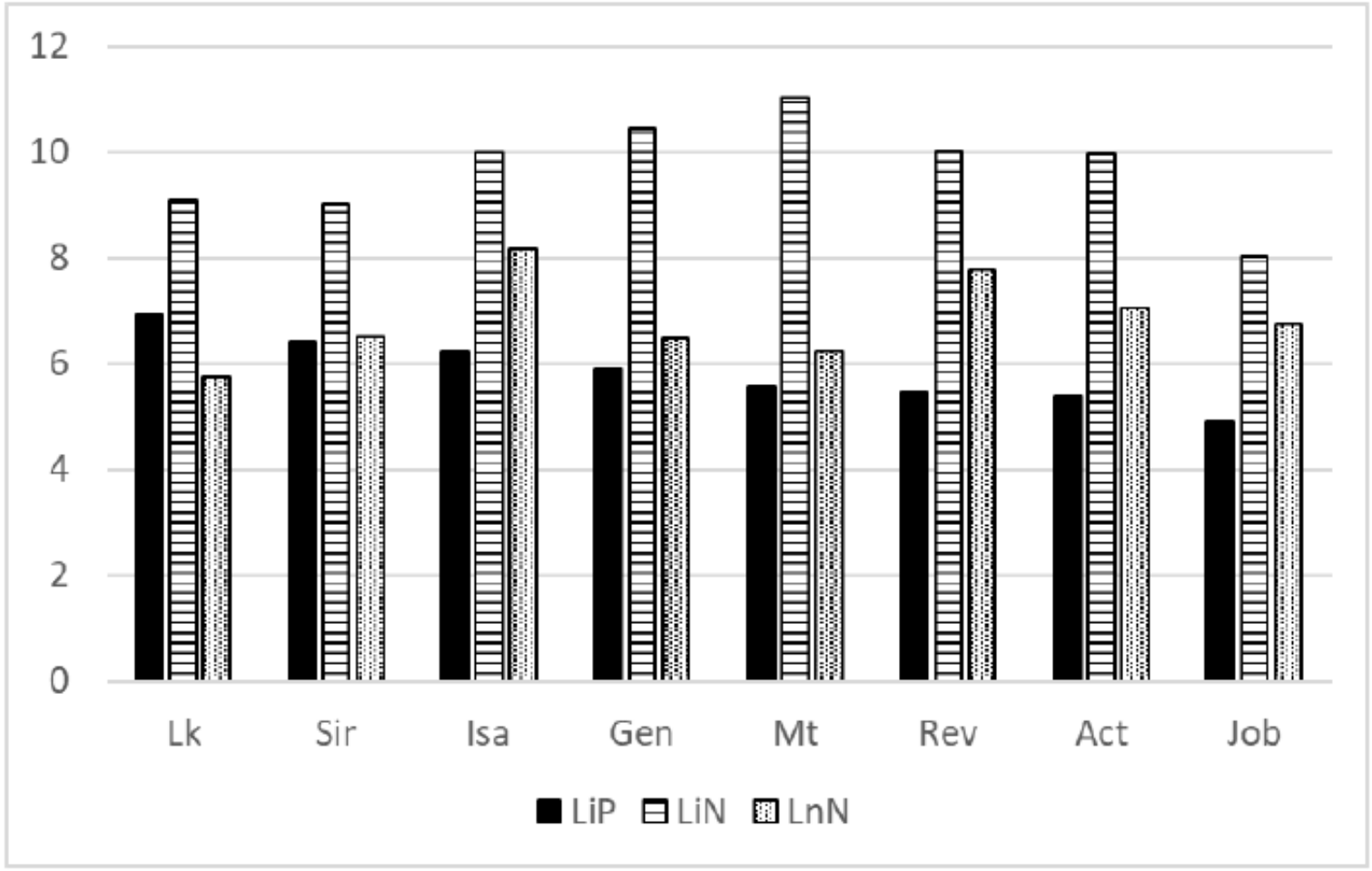
Porovnání délek – jeho interpretace & grafické znázornění



Vztah délky syntaktické fráze a pozice enklitik

- délka fráze měřena v počtu písmen
- enklitika *sě, mi*
- fráze s enklitikem v postiniciální pozici by měla být v průměru kratší než fráze bez enklitika

Vztah délky syntaktické fráze a pozice enklitik



Vztah délky syntaktické fráze a pozice enklitik

	Lk	Sir	Isa	Gen	Mt	Rev	Act	Job	mean	sd
L_iP	6.94	6.41	6.23	5.91	5.58	5.45	5.4	4.9	**5.9**	2.6
L_iN	9.1	9.02	10	10.45	11.01	10.01	9.96	8.02	**10**	6.7
L_nN	5.75	6.52	8.18	6.48	6.23	7.77	7.06	6.74	**6.9**	3.1

Table 10 Average length of analyzed phrases of *sě*

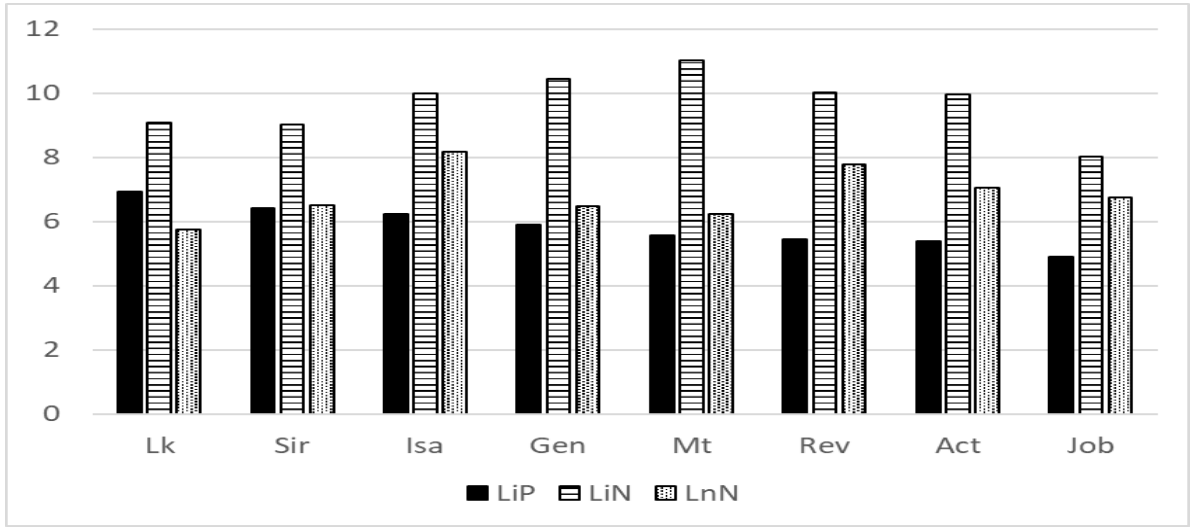


Figure 2 Average length of phrases of *sě* presented in Table 4.

Vztah délky syntaktické fráze a pozice enklitik

	Lk+Sir+Isa+Gen+Mt+Rev+Act+Job	
	mean	**sd**
L_iP	4.82	2.43
L_iN	9.54	6.23
L_nN	6.42	2.04

Table 11 Average length of analyzed phrases of *mi*

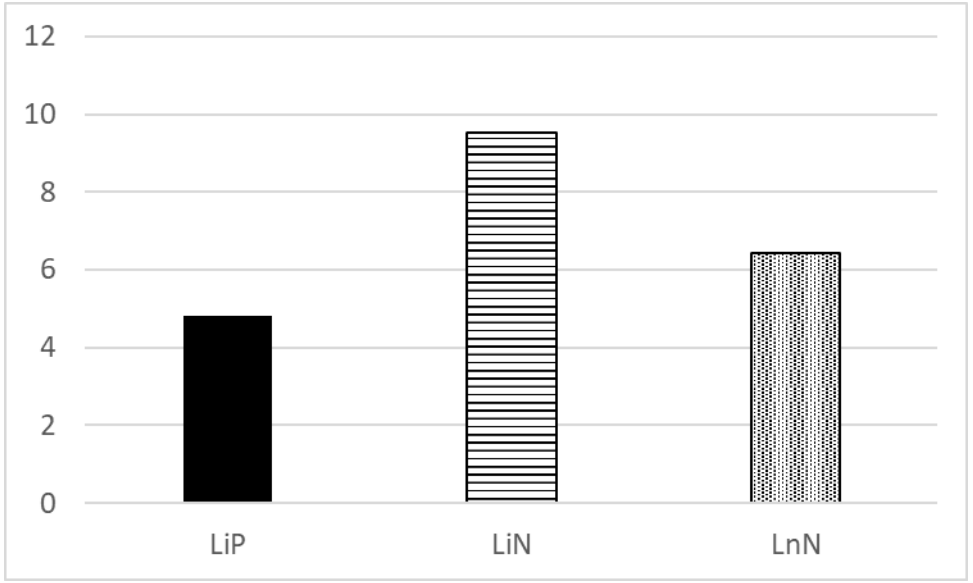
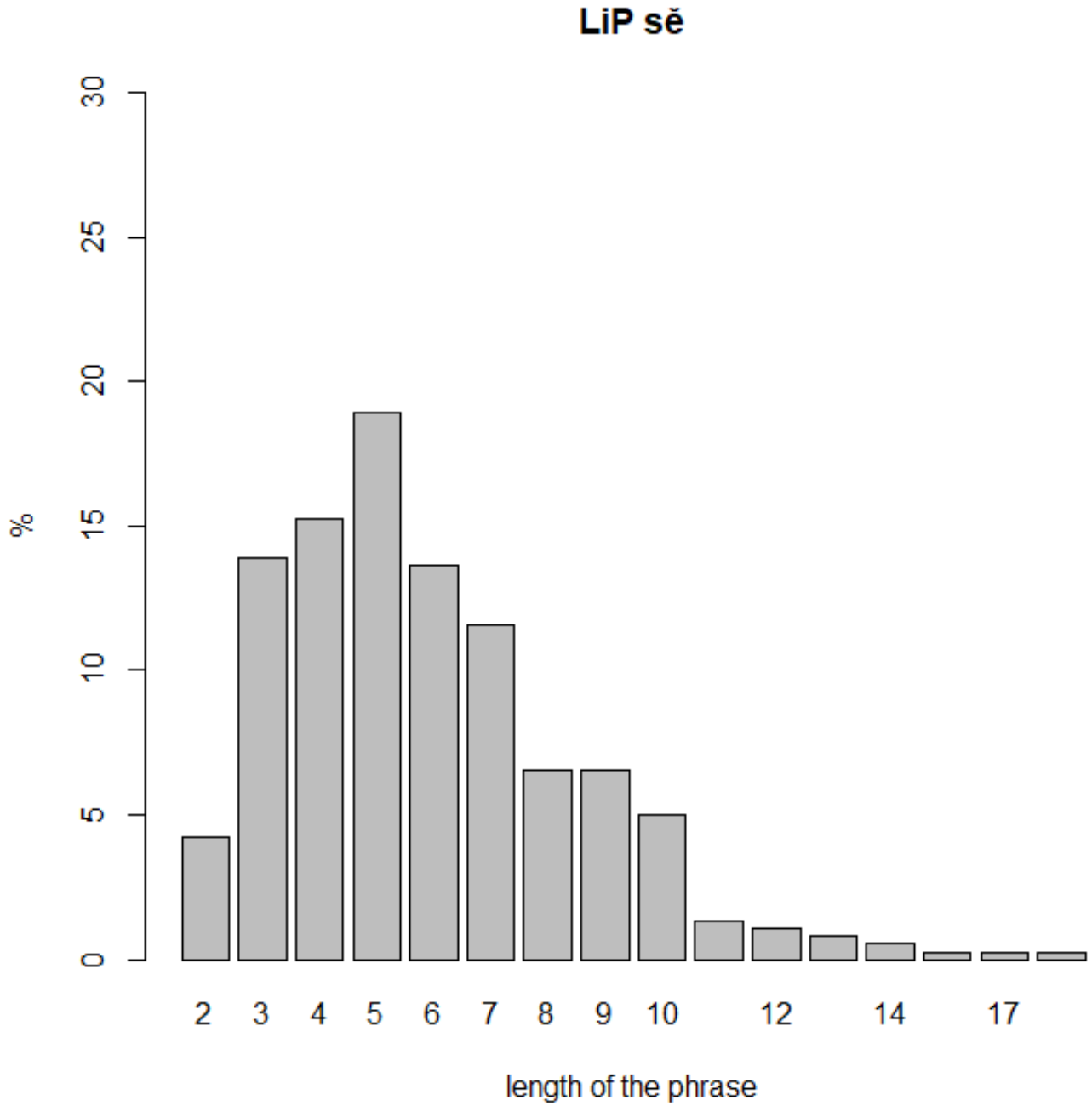
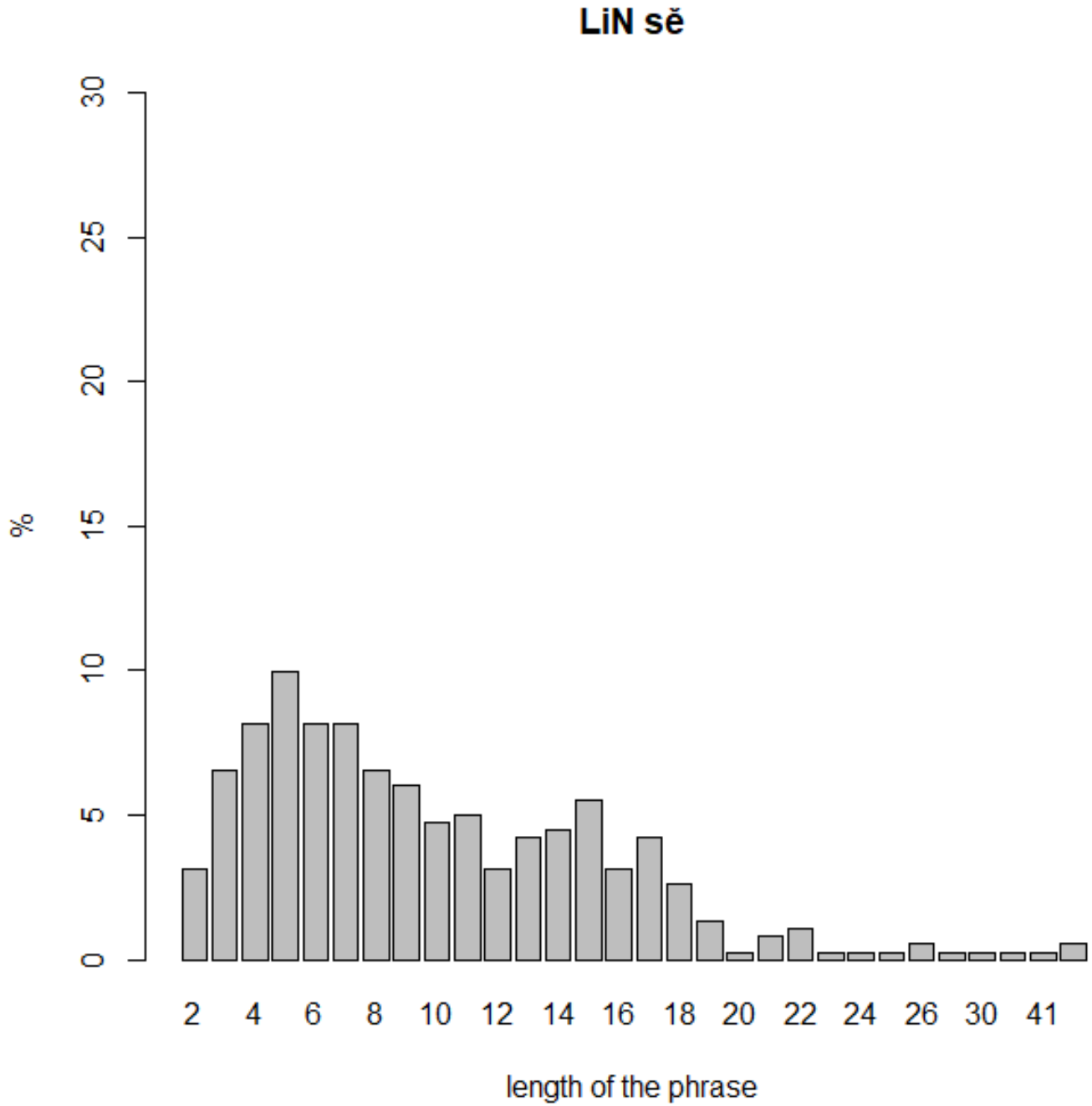
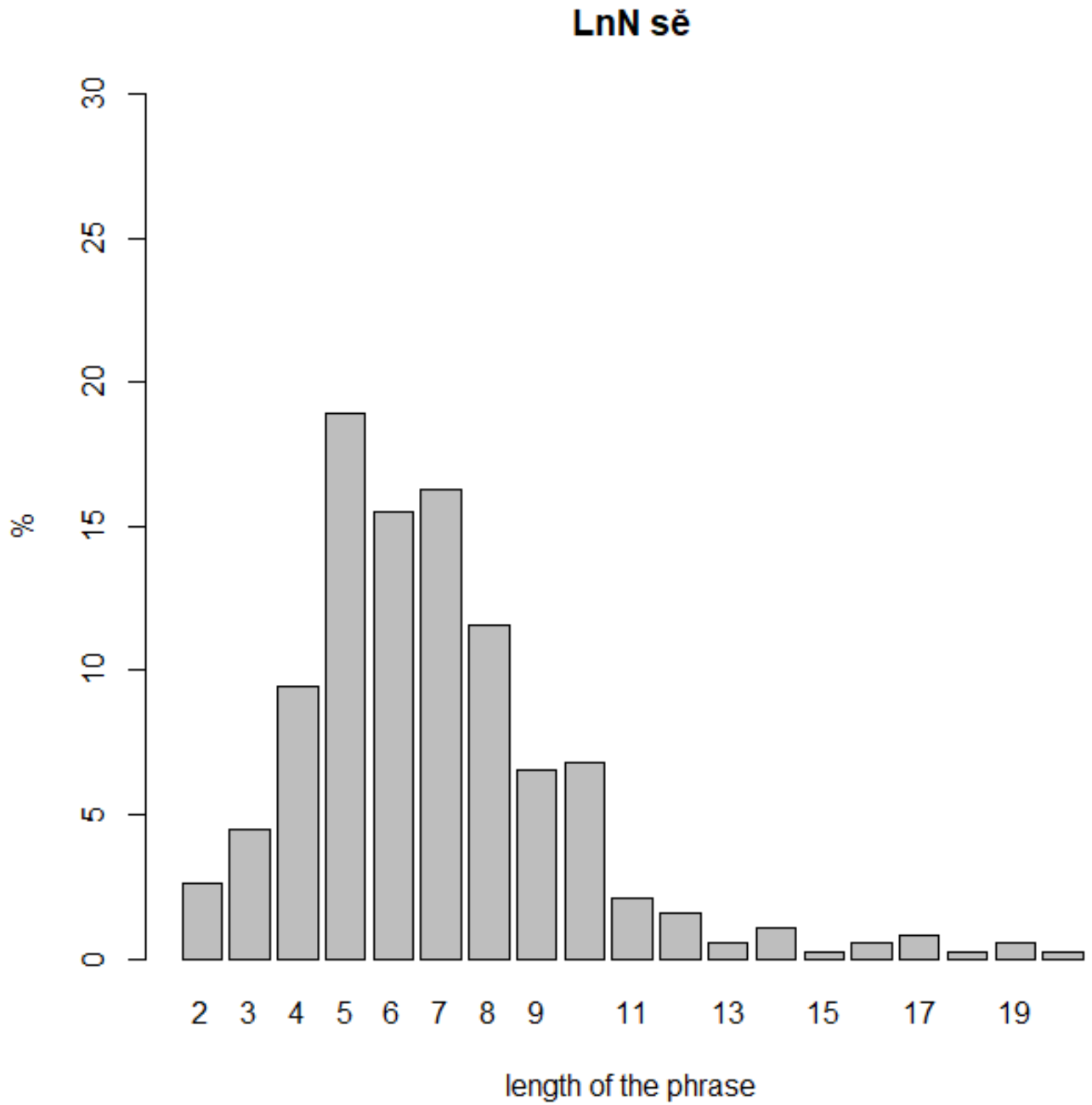


Figure 3 Average length of phrases of *mi* presented in Table 11


LiP sě
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
10
12
14
17
length of the phrase


LiN sě
%
0
5
10
15
20
25
30
2
4
6
8
10
12
14
16
18
20
22
24
26
30
41
length of the phrase


LnN sě
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
11
13
15
17
19
length of the phrase

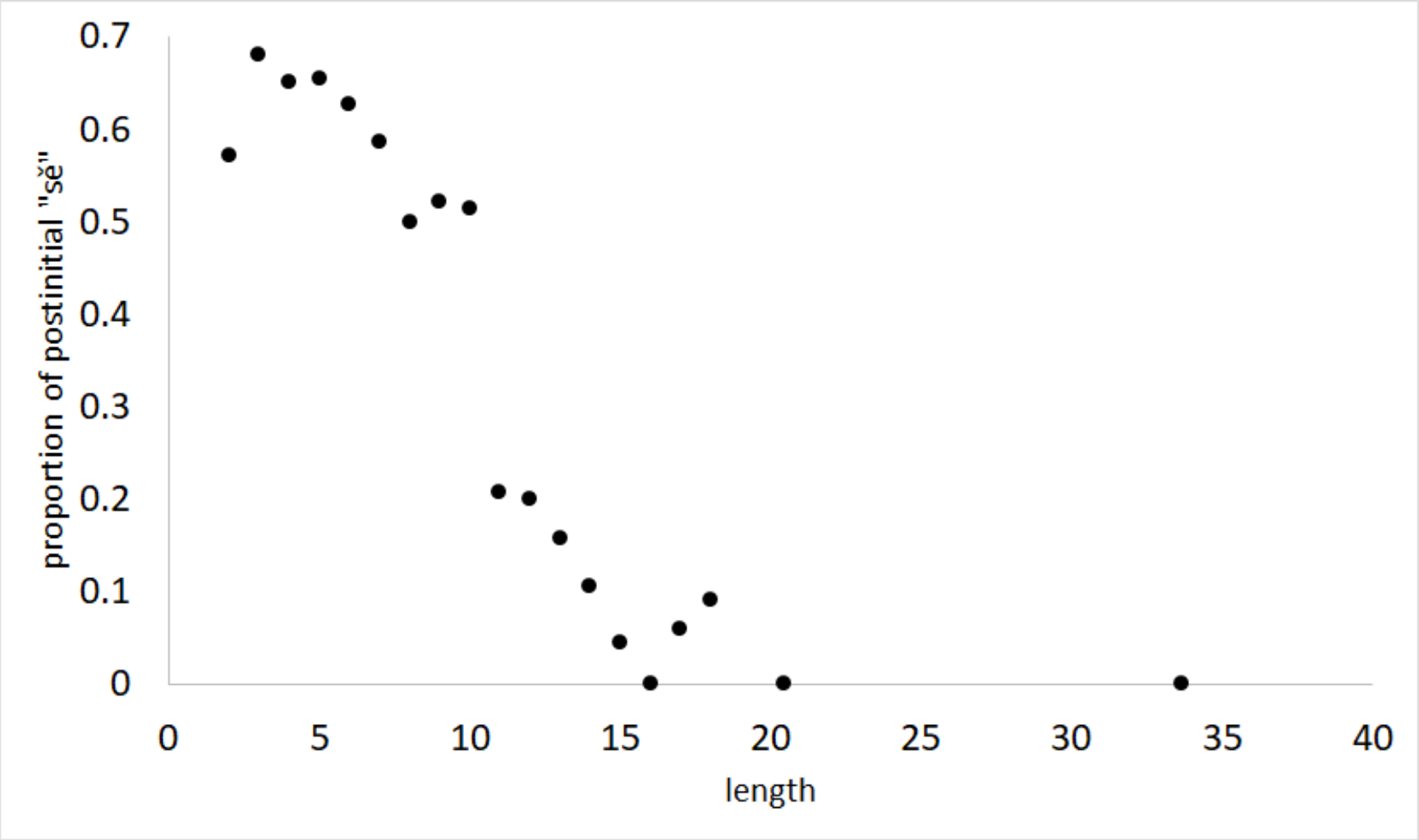
Za hranice popisu…

Za hranice popisu… k testování hypotéz

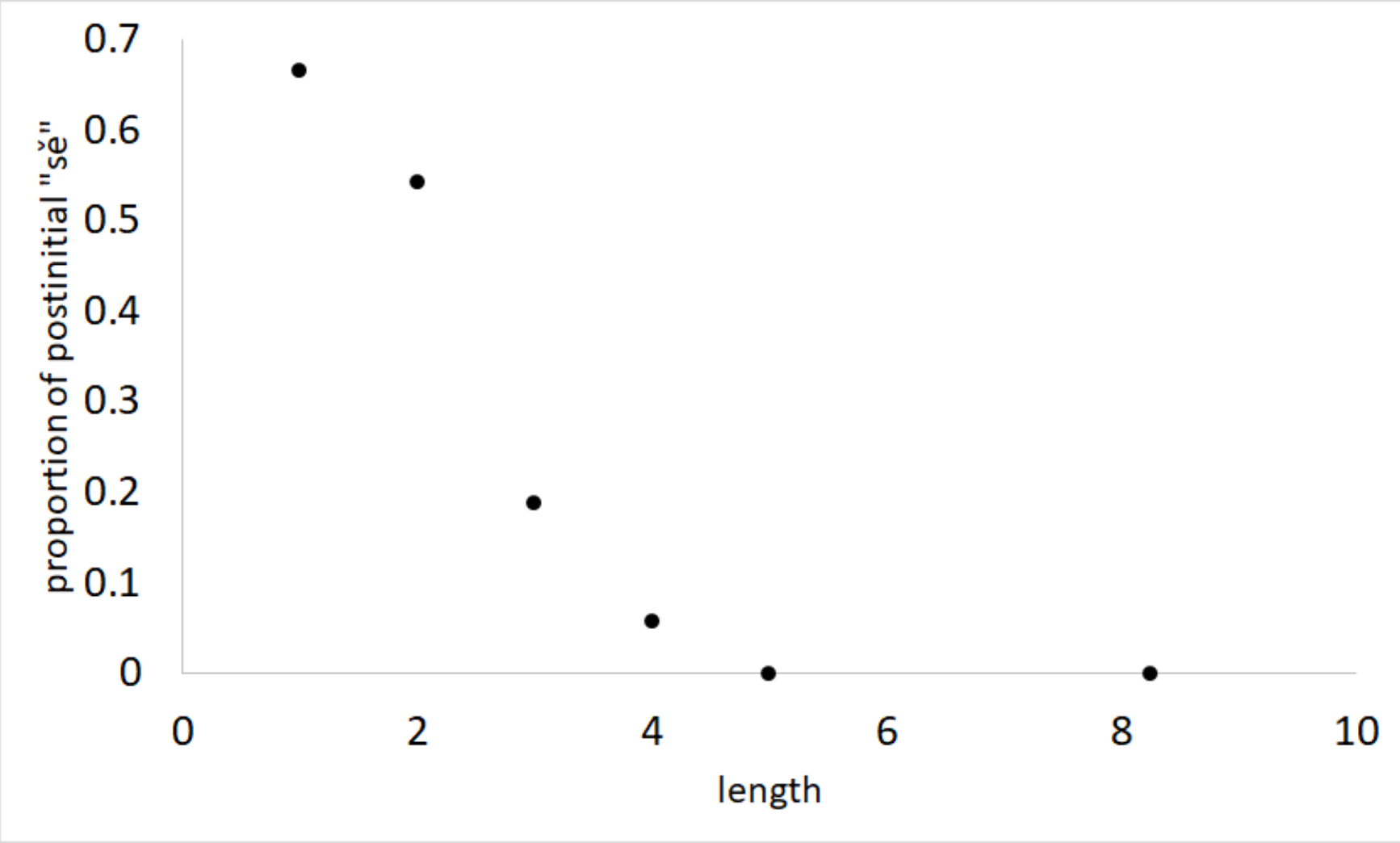
- teoretická zdůvodnění

- hypotéza: čím je iniciální fráze delší, tím menší je pravděpodobnost, že se za ní vyskytne enklitikon

Results - letters



Results - words



Porovnání délek – jeho interpretace

- test…

Za hranice popisu…

- explanace

Za hranice popisu…

- explanace
  - proč má daný systém vlastnosti, jaké pozorujeme

Za hranice popisu…

- explanace
  - proč má daný systém vlastnosti, jaké pozorujeme

- klasifikace & explanace

Teorie & jazykověda

- co je teorie?

Teorie

- A scientific theory is an explanation of an aspect of the natural world that can be repeatedly tested and verified in accordance with the scientific method, using accepted protocols of observation, measurement, and evaluation of results. Where possible, theories are tested under controlled conditions in an experiment. In circumstances not amenable to experimental testing, theories are evaluated through principles of abductive reasoning. Established scientific theories have withstood rigorous scrutiny and embody scientific knowledge.

  (Wikipedia: https://en.wikipedia.org/wiki/Scientific_theory)

Teorie

- A scientific theory is an **explanation** of an aspect of the natural world that can be **repeatedly tested** and verified in accordance with the scientific method, **using accepted protocols of observation, measurement, and evaluation of results**. Where possible, theories are tested under controlled conditions in an experiment. In circumstances not amenable to experimental testing, theories are evaluated through principles of abductive reasoning. Established scientific theories have withstood rigorous scrutiny and embody scientific knowledge.

  (Wikipedia: https://en.wikipedia.org/wiki/Scientific_theory)

Teorie

- The **meaning** of the term scientific theory (often contracted to theory for brevity) as used in the disciplines of science **is significantly different** from the common vernacular usage of theory. In everyday speech, theory can imply an **explanation that represents an unsubstantiated and speculative guess, whereas in science it describes an explanation that has been tested** and widely accepted as valid. These different usages are comparable to the opposing usages of prediction in science versus common speech, where it denotes a mere hope.

  (Wikipedia: https://en.wikipedia.org/wiki/Scientific_theory)

Teorie

- lingvistické teorie?

Hypotéza

- Co je hypotéza?

Hypotéza

- Co je hypotéza?
- Formální vlastnosti hypotézy?

Hypotéza

- Co je hypotéza?
- Formální vlastnosti hypotézy?
- Lingvistické hypotézy…

„Hypotéza“ v lingvistice

- „V hláskosloví ani v jiných rovinách vodňanského herbáře nejsou prokazatelné další nářeční jevy z oblasti, kde nedošlo ke vzniku vibranty *ř* (východomoravské území), dáváme proto přednost **hypotéze**, že se jedná o nedbalý zápis“ (Černá 2005, s. 76);
- „V nich se překlad Františka Vrby jeví jako silně zatížený mužským genderovým úhlem pohledu a estetikou vnímání; potvrzuje se tak původní **hypotéza**, že se spíše „staví na stranu“ mužského hrdiny, resp. autorského tvůrce a erotické líčení prezentuje spíše z jeho perspektivy…“ (Širokovská 2004, s. 23);
- „Proč není samo slovo *plémě* ve staročeských textech doloženo v očekávaném významu, o tom lze vznášet různé **hypotézy**.“ (Šimandl 2007, s. 238);
- „**Hypotéza** 2.1: Co-text je věrným zrcadlem (situačního) kontextu v tom smyslu, že všechny pro danou komunikační situaci relevantní kontextové vlastnosti jsou co-textem explicitně reflektovány, a mají tedy nějaký textový korelát. (…) Hypotéza 2.2: (Textový) kontext věrně reflektuje všechny vlastnosti jazykových jevů relevantní pro jejich užití. (Cvrček 2013, s. 24)“;
- „Vycházeje z toho, že teorie valence i přes zjevná slabá místa představuje dobrý konstrukt lingvistické teorie, pokusím se nyní představit **hypotézu** *modifikované valenční teorie* (MVT) a formulovat základní principy této teorie.“ (Karlík 2001, s. 171n).

Empiricky testovatelná hypotéza

- předpokládaný vztah mezi dvěma vlastnostmi = působení mechanismu

Empiricky testovatelná hypotéza

- předpokládaný vztah mezi dvěma vlastnostmi = působení mechanismu
- teoretické zdůvodnění

Hypotéza (Greis 2009, s. 11)

- tvrzení, které se týká více než jednoho jevu či případu;

Hypotéza (Greis 2009, s. 11)

- tvrzení, které se týká více než jednoho jevu či případu;

- má alespoň implicitně strukturu podmínkového souvětí, tj. *„jestliže…, pak…“*, případně *„čím…, tím…“* (např. čím je slovo frekventovanější, tím je kratší);

Hypotéza (Greis 2009, s. 11)

- tvrzení, které se týká více než jednoho jevu či případu;
- má alespoň implicitně strukturu podmínkového souvětí, tj. *„jestliže…, pak…“*, případně *„čím…, tím…“* (např. čím je slovo frekventovanější, tím je kratší);
- je falzifikovatelné (tj. vyvratitelné) prostřednictvím experimentu, který dovoluje rozhodnout, zda predikce formulovaná prostřednictvím hypotézy je vyvrácena, či ne
  - (vyhodnocení experimentu většinou pomocí statistických testů).

Hypotéza

- která tvrzení *jsou/nejsou* testovatelnými hypotézami?
  1. *hodně mužů má pleš*
  2. *pokud se v knize vyskytují biblický příběh, je to apokryf*
  3. *jestli se zavedou řidičáky „na zkoušku“, může se snížit nehodovost mladých řidičů a řidiček*
  4. *muži mají častěji pleš než ženy*
  5. *jestliže se je sloveso dokonavé, častěji se na něj váže přímý akuzativní předmět než na sloveso nedokonavé*
  6. *ženy jsou citlivé*
  7. *čím je slovo frekventovanější, tím je větší jeho polysémie*
  8. *jestli se zavedou řidičáky „na zkoušku“, sníží se nehodovost mladých řidičů a řidiček*
  9. *nářečí často ovlivňují podobu mluveného jazyka obyvatel dané nářeční oblasti*

Hypotéza

- Wikipedie
  - https://cs.wikipedia.org/wiki/Hypot%C3%A9za

Hypotéza - opakování

- která tvrzení *jsou/nejsou* testovatelnými hypotézami?
  1. *delší klauze (měřeno vpočtu slov) mají v průměru kratší slova (měřeno v počtu slabik) než klauze kratší*
  2. *v odborných textech je hodně dlouhých vět*
  3. *pokud je slovo syntakticky závislé na substantivu, je to přívlastek*
  4. *auxiliáry jsou v průměru kratší něž autosémantika*
  5. *mezi délkou slova měřenou v počtu hlásek a v počtu slabik je lineární závislost*
  6. *děti z měst mají bohatou slovní zásobu*
  7. *čeština je jeden z nejkomplikovanějších jazyků na světě*
  8. *čím je slovo delší, tím má více hlásek*
  9. *čím je člověk starší, tím v průměru používá více zájmen*

Populace & vzorek

- populace – základní soubor
  - úplná množina prvků

Populace & vzorek

- populace – základní soubor
  - úplná množina prvků

- co je v jazyce „základním souborem“?

Populace & vzorek

- populace – základní soubor
  - úplná množina prvků

- co je v jazyce „základním souborem“?
  - otázka reprezentativnosti…

Populace & vzorek

- vzorek – výběrový soubor
  - výběr ze základního souboru

Populace & vzorek

- vzorek – výběrový soubor
  - výběr ze základního souboru

- ze vzorku je možné vyvozovat závěry pro celou populaci
  - statistické testy
  - rozdíly, náhoda

Statistické testy významnosti

- porovnávají se dvě hypotézy
  - **nulová hypotéza**:
    tvrzení, které obvykle deklaruje “žádný rozdíl”, tj. nalezený rozdíl je dán variablitou dat, náhodou
    - (např. mince není falešná; mezi formou jazyka a četností užívaní *bychom*/*bysme* není rozdíl)

Statistické testy významnosti

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

Statistické testy významnosti

- testuje se platnost H0

- hladina významnosti
  - pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv ona platí
  - obvykle 5 % (0,05) nebo 1 % (0,01)
  - p-hodnota (p-value)

Statistické testy významnosti

- hladina významnosti
  - pravděpodobnost, že se zamítne nulová hypotéza, ačkoliv ona platí
  - obvykle 5 % (0,05) nebo 1 % (0,01)

- konvence
  - chyba 1. typu (neadekvátní zamítnutí H0, odpovídá hladině významnosti)
  - chyba 2. typu (neadekvátní nezamítnutí H0)

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?
  - 59x panna, 41 orel → podvádí se?

Statistické testy významnosti

- hod mincí (100x)
  - 50x panna a 50x orel → podvádí se?
  - 52x panna a 48x orel → podvádí se?
  - 98x panna, 2x orel → podvádí se?
  - 59x panna, 41 orel → podvádí se?
  - 60x panna, 40 orel → podvádí se?
  - …

Statistické testy významnosti

- pokud padne panna 61x, tak je větší než 95% pravděpodobnost, že jeden z hráčů podvádí
- jinými slovy: pravděpodobnost, že budeme neoprávněně tvrdit, že jeden z hráčů nepodvádí, je menší než 5%

Statistické testy významnosti

- testuje se platnost H0

Statistické testy významnosti

- testuje se platnost H0
- odmítnutí H0 **neznamená, že H1 platí**

Statistické testy významnosti

- testuje se platnost H0
- odmítnutí H0 **neznamená, že H1 platí**
- odmítnutí H0 **znamená, že existuje určitá/vysoká pravděpodobnost toho, že naměřený rozdíl není možné vysvětlit vlivem náhody**
- H1 se nikdy **nepotvrzuje** (confirmation), vždy se jedná o **vyvracení (rejection)** H0 nebo H1
  - terminologická poznámka: QL → corroboration

Chí-kvadrát test dobré shody

- příklad: předpokládáme, že v románech se bude častěji používat nespisovná varianta slova “bychom” než v publicistických textech
  - proměnnými jsou: a) typ textu; b) varianta slova

  H_0: mezi typem textu a používáním nespisovné varianty slova “bychom” není žádný vztah

  H_1: mezi typem textu a používáním nespisovné varianty slova “bychom” je vztah, tj. tato forma se častěji vyskytuje v próze

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6
p = 0,00000000000000022		

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6
p = 0,00000000000000022		

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2
p = 0,0000001851		

Chí-kvadrát test dobré shody

	SYN2005nov (romány)	SYN2005pub (publicistika)
bychom	5260	6679
bysme	714	39
% bysme	13,6	0,6
$p < 0{,}001$		

	SYN2005nov (romány)	SYN2005col (povídky)
bychom	5260	1660
bysme	714	136
% bysme	13,6	8,2
$p < 0{,}001$		

Příklad – hypotéza tranzitivity

- Hopper, P., Thompson, S. (1980). Transitivity in Grammar and Discourse. Language 56, 251-299.

Table 1: Transitivity parameters

		high T	low T
A	PARTICIPANTS	2 or more	1
B	KINESIS	action	non-action
C	ASPECT	telic	atelic
D	PUNCTUALITY	punctual	non-punctual
E	VOLITIONALITY	volitional	non-volitional
F	AFFIRMATION	affirmative	negative
G	MODE	realis	irrealis
H	AGENCY	A high in potency	A low in potency
I	AFFECTEDNESS of O	O totally affected	O not affected
J	INDIVIDUATION of O	O highly individuated	O non-individuated

Hypotéza tranzitivity

- “[t]ransitivity is a crucial relationship in language, having a number of universally predictable consequences in grammar”
- transitivity “can be broken into its component parts (…), they allow clauses to be characterized as MORE or LESS Transitive: the more features a clause has in the 'high' column in 1A–J, the more Transitive it is”

Hypotéza tranzitivity

- “If two clauses (a) and (b) in a language differ in that (a) is higher in Transitivity according to any features 1A-J, then, if concomitant grammatical or semantic difference appears elsewhere in the clause, that difference will also show (a) to be higher in Transitivity”
- “whenever two values of the transitivity components are necessarily present (...) they will agree in being either both high or both low in value”.
- The co-variation has to be viewed not in the strict sense, but as a tendency.

Hypotéza tranzitivity

	perfective verb	imperfective verb	percentage of proper names objects
proper name object	382	270	58.6%
common noun object	5255	5878	47.2%

$\chi^2 = 32.01$ p-hodnota < 0.05

	affirmative predicate	negative predicate	percentage of proper names objects
proper name object	744	48	93.9%
common noun object	13794	1215	91.9%

$\chi^2 = 4.23$ p-hodnota < 0.05

Hypotéza tranzitivity

	1 participant clauses	2 or more participant clauses	% 1 participant clauses
PSC	292	2054	12.5%
PDT	230	1348	14.6%

$\chi_1^2 = 3.71$, $P = 0.0542$

	imperfective predicates	perfective predicates	% imperfective predicates
PSC	245	44	84.8%
PDT	158	72	68.7%

$\chi_1^2 = 19.08$, $P = 0.00001$

Statistické testy

- četnosti
- průměry
- korelace

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

- Np_i … očekávané četnosti
- X_i … naměřené četnosti

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	žánr C	**žánr D**	**Σ**
slovo A	X_1	X_3	X_1+X_3
slovo B	X_2	X_4	X_2+X_4
Σ	X_1+X_2	X_3+X_4	$X_1+X_2+X3+X4$
	Np_1	Np_3	
	Np_2	Np_4	

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	žánr C	žánr D	Σ
slovo A	X_1	X_3	X_1+X_3
slovo B	X_2	X_4	X_2+X_4
Σ	X_1+X_2	X_2+X_4	$X_1+X_2+X3+X4$
	Np_1	Np_3	
	Np_2	Np_4	

$$N_{p1} = \frac{(x_1 + x_3) \cdot (x_1 + x_2)}{x_1}$$

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	žánr C	**žánr D**	**Σ**
slovo A	X_1	X_3	X_1+X_3
slovo B	X_2	X_4	X_2+X_4
Σ	X_1+X_2	X_2+X_4	$X_1+X_2+X3+X4$
	Np_1	Np_3	
	Np_2	Np_4	

	žánr C	**žánr D**	**Σ**
slovo A	10	10	20
slovo B	20	20	40
Σ	30	30	60
	10,00	10,00	
	20,00	20,00	

$\chi^2 = 0$, p-hodnota = 1

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	žánr C	**žánr D**	**Σ**
slovo A	X_1	X_3	X_1+X_3
slovo B	X_2	X_4	X_2+X_4
Σ	X_1+X_2	X_2+X_4	$X_1+X_2+X3+X4$
	Np_1	Np_3	
	Np_2	Np_4	

	žánr C	**žánr D**	**Σ**
slovo A	5	10	15
slovo B	25	20	45
Σ	30	30	60
	7,50	7,50	
	22,50	22,50	

$\chi^2 = 1{,}42$, p-hodnota = 0,23

Test dobré shody chi-kvadrát

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

	žánr C	**žánr D**	**Σ**
slovo A	X_1	X_3	X_1+X_3
slovo B	X_2	X_4	X_2+X_4
Σ	X_1+X_2	X_2+X_4	$X_1+X_2+X2+X4$
	Np_1	Np_3	
	Np_2	Np_4	

	žánr C	**žánr D**	**Σ**
slovo A	5	20	25
slovo B	25	20	45
Σ	30	40	70
	10,71	14,29	
	19,29	25,71	

$\chi^2 = 6{,}91$, p-hodnota = 0,004

Test dobré shody chi-kvadrát

- Excel
- vypočítat očekávané hodnoty
- pak CHISQ.TEST

Test dobré shody chi-kvadrát

- otestujte hypotézu závislosti výskytu daných slov na žánru

	žánr C	žánr D	žánr E
slovo A	5	20	18
slovo B	25	20	26

Test dobré shody chi-kvadrát

- post hoc test

	žánr C	žánr D	žánr E
slovo A	5	20	18
slovo B	25	20	26

Test dobré shody chi-kvadrát

- Wikipedia
  - https://cs.wikipedia.org/wiki/Test_dobr%C3%A9_shody

- Čech, R., Pajas, P. (2009). Pitfalls of the Transitivity Hypothesis: Transitivity in Conversation and Written Language in Czech. Glottotheory 2, 41-49.

Test dobré shody chi-kvadrát

- omezení
  - malé počty: očekávané četnosti > 5
  - nevhodný pro velká data

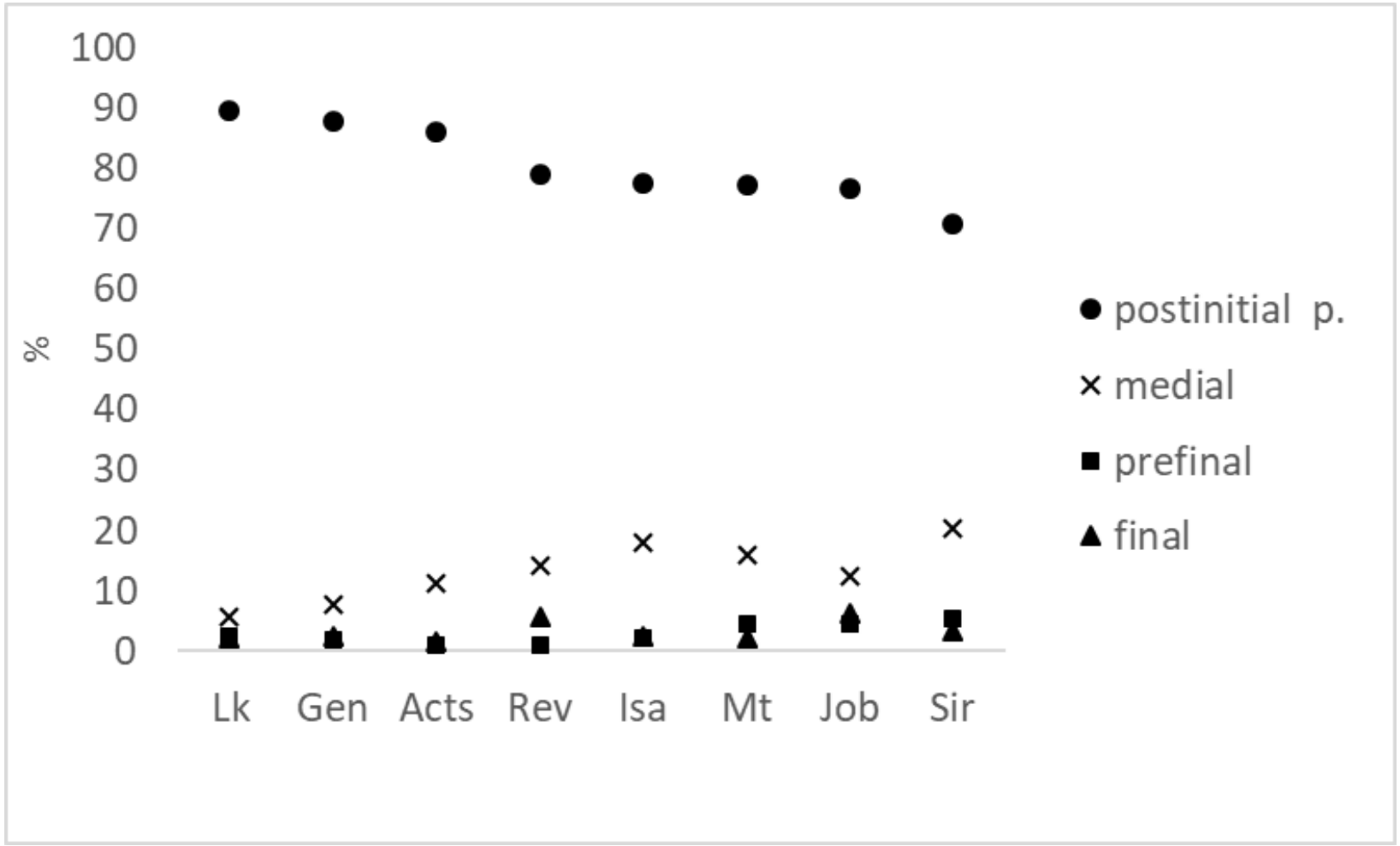
	romány	novely	Σ	% novely
konstrukce *A*	500000	501800	1001800	50,09%
konstrukce *B*	501500	500000	1001500	49,93%
Σ	1001500	1001800	2003300	
chi^2 = 5.43, p=0,020				

Příklad: vliv typu textu (žánru) na postavení enklitik

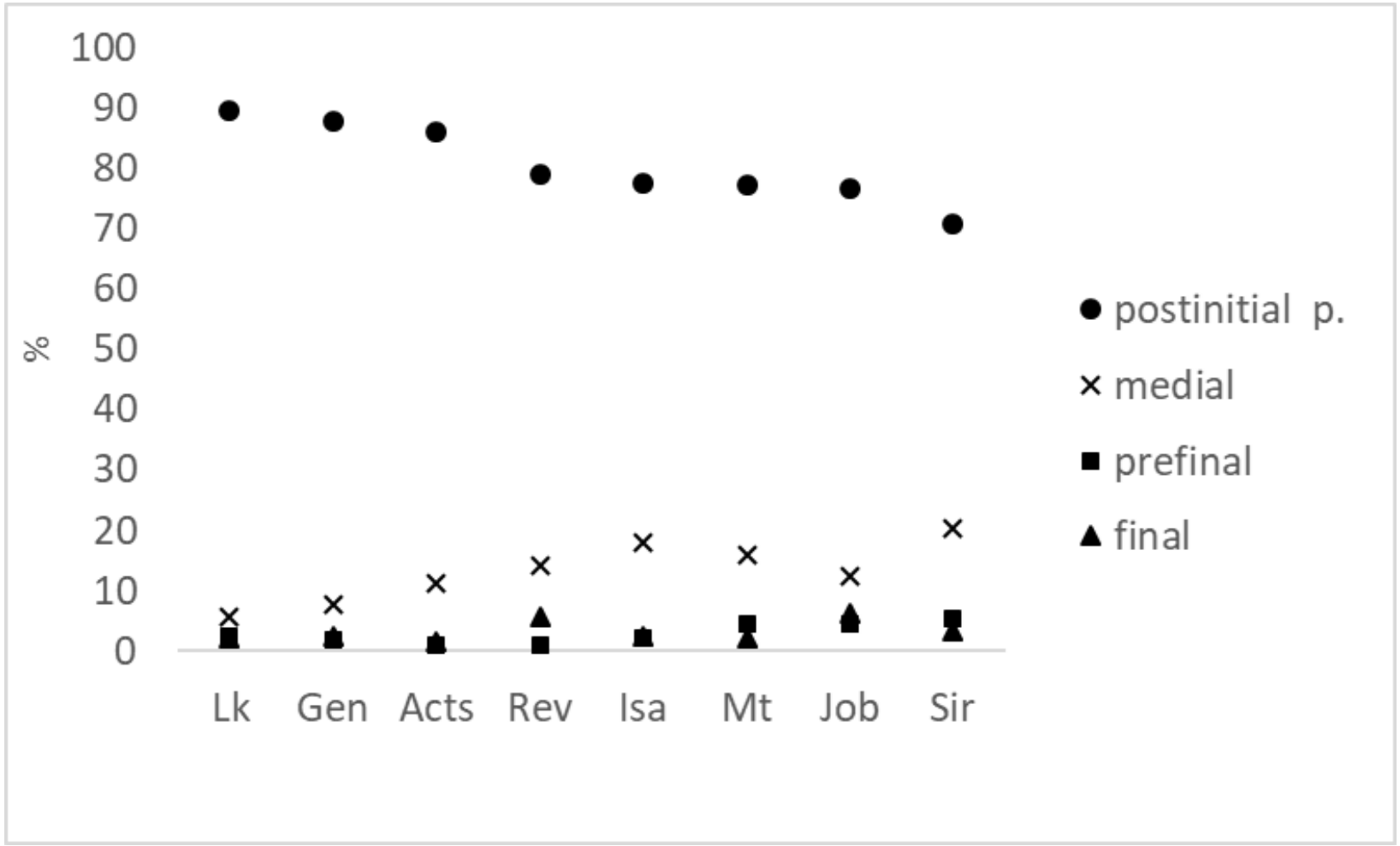
- H0: typ textu nemá vliv na postavení enklitik
- H1: typ textu má vliv na postavení enklitik

Kosek, P., Navrátilová, O., Čech, R., Mačutek, J. (2018). Word Order of Reflexive 'sě' in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 2). Studia Linguistica Universitatis Iagellonicae Cracoviensis, 135, 3, 189-200.
http://www.cechradek.cz/publ/2018_Kosek_etal_Krakow_j_02.pdf

Příklad: vliv typu textu (žánru) na postavení enklitik



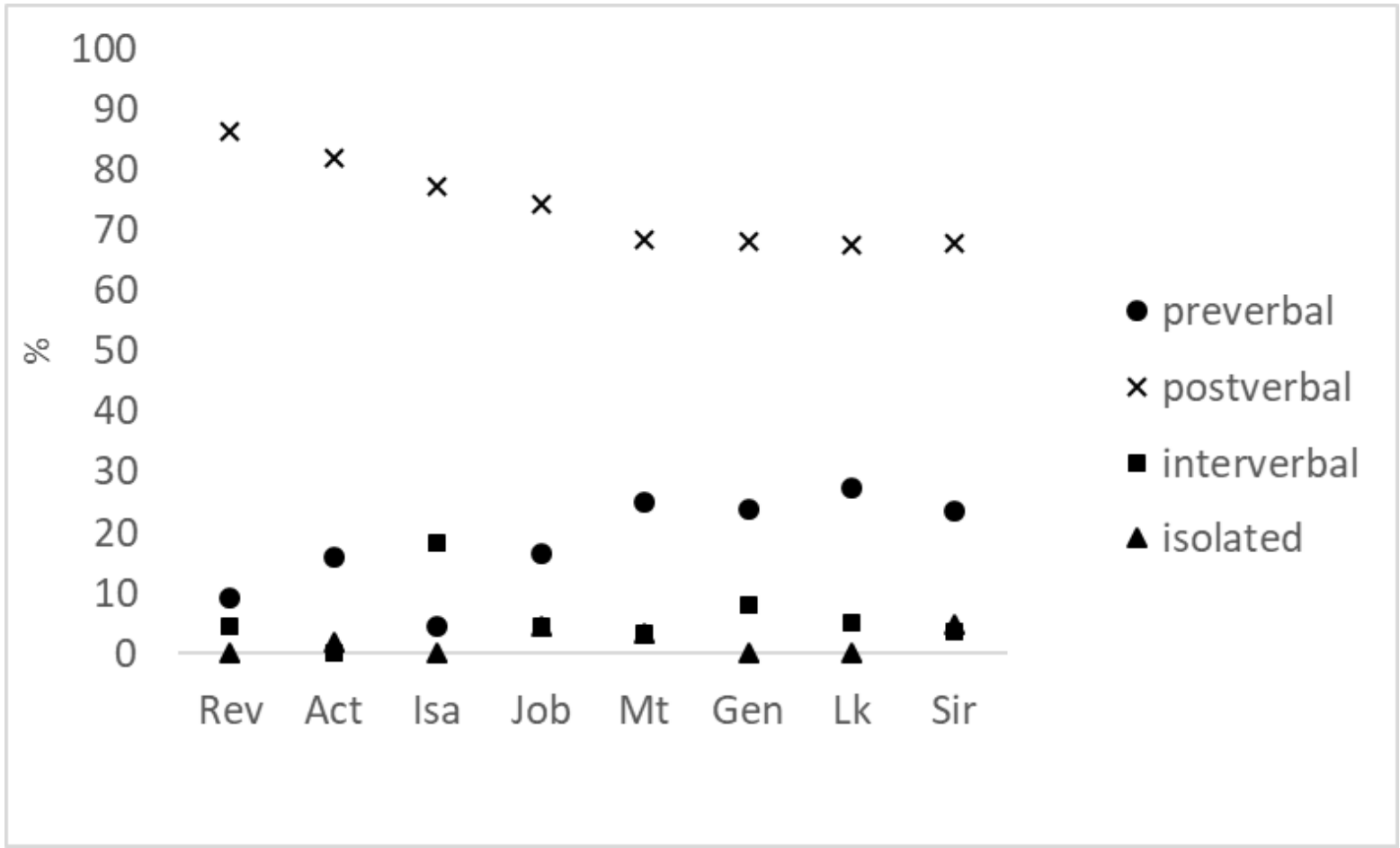
Příklad: vliv typu textu (žánru) na postavení enklitik



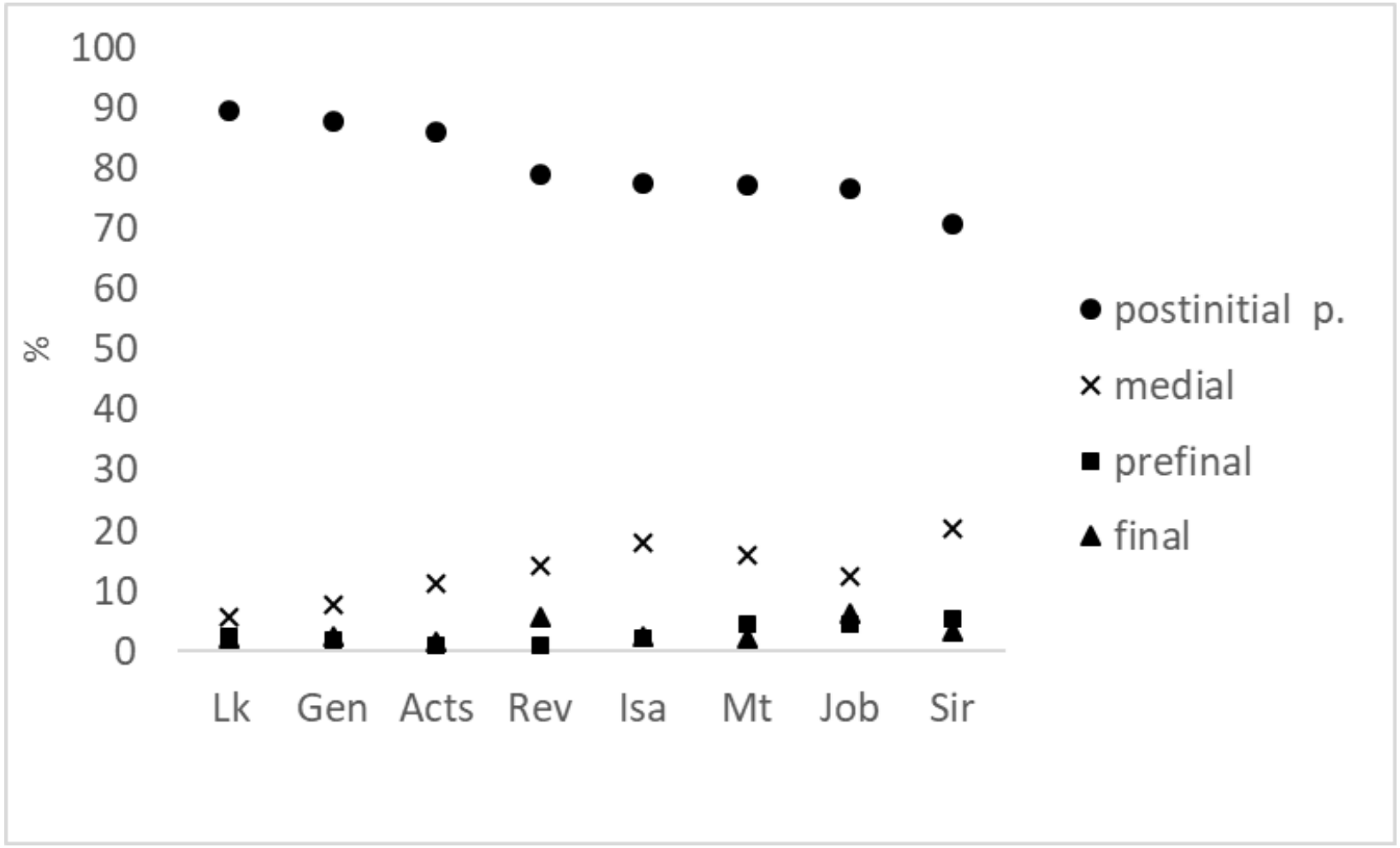
$\chi^2 = 83.712$

p-value < 0.001

Příklad: vliv typu textu (žánru) na postavení enklitik



Příklad: vliv typu textu (žánru) na postavení enklitik



$\chi^2 = 33.772$

p-value < 0.03

Test dobré shody chi-kvadrát

- jak spočítat
  - manuálně
  - Excel – viz návody
  - online nástroje
    - např. https://www.socscistatistics.com/tests/
  - R software
    - https://cran.r-project.org/

Úkol

H0: mezi četnostmi výrazů děkuji a děkuju a typem textu není vztah
H1: mezi četnostmi výrazů děkuji a děkuju a typem textu je vztah

materiál: SYN2020
typy textů: FIC: beletrie, NMG: publicistika, NFC: oborová literatura

intuice?

zjistěte hodnoty z ČNK

První pohled?

	děkuji	děkuju
FIC: beletrie	2345	1936
NMG: publicistika	640	130
NFC: oborová literatura	582	115

První pohled?

	děkuji	děkuju
FIC: beletrie	2345	1936
NMG: publicistika	640	130
NFC: oborová literatura	582	115

vypočítejte procentuální zastoupení děkuji v jednotlivých typech textu

Druhý pohled?

	děkuji	děkuju	% děkuji
FIC: beletrie	2345	1936	54.78 %
NMG: publicistika	640	130	83.12 %
NFC: oborová literatura	582	115	83.5 %

vytvořte tabulku, v níž budou očekávané četnosti, použijte Excel

Druhý pohled?

	děkuji	děkuju	% děkuji
FIC: beletrie	2345	1936	54.78 %
NMG: publicistika	640	130	83.12 %
NFC: oborová literatura	582	115	83.5 %

Očekávané frekvence

pozorované			
	děkuji	děkuju	suma
FIC: beletrie	2345	1936	4281
NMG: publicistika	640	130	770
NFC: oborová literatura	582	115	697
suma	3567	2181	5748
očekávané			
	děkuji	děkuju	suma
FIC: beletrie	2656.63	1624.37	4281
NMG: publicistika	477.83	292.17	770
NFC: oborová literatura	432.53	264.47	697
suma	3567	2181	5748

Test

- https://www.socscistatistics.com/tests/chisquare2/default2.aspx

Results						
	dekuji	dekuju				***Row Totals***
FIC	2345 (2656.63) [36.56]	1936 (1624.37) [59.79]				4281
NMG	640 (477.83) [55.04]	130 (292.17) [90.01]				770
NFC	582 (432.53) [51.65]	115 (264.47) [84.47]				697
Column Totals	3567	2181				**5748 (Grand Total)**

The chi-square statistic is 377.511. The *p*-value is < 0.00001. The result is significant at $p < .05$.

Cvičení

- data:

Table 7. The distribution of imperfective and perfective predicates in one and two or more participant clauses in the PSC

	imperfective predicates	perfective predicates	% imperfective predicates
1 participant clauses	158	72	68.7%
2 or more participant clauses	887	460	65.9%

- https://www.socscistatistics.com/tests/chisquare/default2.aspx

Opakování

- co znamená aplikace statistického testu?
- jaké závěry lze vyvodit z aplikace statistického testu?
- jaký je vztah statistického testuj s ohledem na populaci a vzorek?

Opakování

- vyhodnoťte vztah mezi perfektivitou a mono/ditranzitivitou slovesa
- hypotéza: perfektivní slovesa by se měla častěji realizovat jako ditranzitivní než monotranzitivní
- náležitě interpretujte výsledky

PDT		ditrnas.	monotrans.	% ditrans
doporučit	perf.	31	23	
doporučovat	imperf.	18	38	
poskytnout	perf.	28	23	
poskytovat	imperf.	21	37	

- https://www.socscistatistics.com/tests/chisquare/