

# Úvod do kvantitativní lingvistiky

ZS 2023

# Statistické testy významnosti

- co znamená statistický test?

# 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 variabilitou dat, náhodou
    - (např. mince není falešná; mezi formou jazyka a četností užívání *bychom/bysme* není rozdíl)



# 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

# Statistické testy významnosti

- testuje se platnost  $H_0$
- 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  $H_0$ , odpovídá hladině významnosti)
  - chyba 2. typu (neadekvátní nezamítnutí  $H_0$ )

# 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  $H_0$

# Statistické testy významnosti

- testuje se platnost  $H_0$
- odmítnutí  $H_0$  **neznamená, že  $H_1$  platí**

# Statistické testy významnosti

- testuje se platnost  $H_0$
- odmítnutí  $H_0$  **ne**znamená, že  $H_1$  platí
- odmítnutí  $H_0$  **znamená**, že **existuje určitá/vysoká pravděpodobnost toho, že naměřený rozdíl není možné vysvětlit vlivem náhody**
- $H_1$  se nikdy **nepotvrzuje** (confirmation), vždy se jedná o **vyvracení (rejection)**  $H_0$  nebo  $H_1$ 
  - terminologická poznámka: QL → corroboration

# 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	$\Sigma$
<b>slovo A</b>	$X_1$	$X_3$	$X_1+X_3$
<b>slovo B</b>	$X_2$	$X_4$	$X_2+X_4$
<b><math>\Sigma</math></b>	$X_1+X_2$	$X_3+X_4$	$X_1+X_2+X_3+X_4$
	$Np_1$	$Np_3$	
	$Np_2$	$Np_4$	

# Test dobré shody chi-kvadrát

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

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	$X_1$	$X_3$	$X_1+X_3$
<b>slovo B</b>	$X_2$	$X_4$	$X_2+X_4$
<b><math>\Sigma</math></b>	$X_1+X_2$	$X_2+X_4$	$X_1+X_2+X_2+X_4$
	$Np_1$	$Np_3$	
	$Np_2$	$Np_4$	

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



# Test dobré shody chi-kvadrát

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

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	$X_1$	$X_3$	$X_1+X_3$
<b>slovo B</b>	$X_2$	$X_4$	$X_2+X_4$
<b><math>\Sigma</math></b>	$X_1+X_2$	$X_3+X_4$	$X_1+X_2+X_3+X_4$
	$Np_1$	$Np_3$	
	$Np_2$	$Np_4$	

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	10	10	20
<b>slovo B</b>	20	20	40
<b><math>\Sigma</math></b>	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	$\Sigma$
<b>slovo A</b>	$X_1$	$X_3$	$X_1+X_3$
<b>slovo B</b>	$X_2$	$X_4$	$X_2+X_4$
<b><math>\Sigma</math></b>	$X_1+X_2$	$X_3+X_4$	$X_1+X_2+X_3+X_4$
	$Np_1$	$Np_3$	
	$Np_2$	$Np_4$	

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	5	10	15
<b>slovo B</b>	25	20	45
<b><math>\Sigma</math></b>	30	30	60
	7,50	7,50	
	22,50	22,50	

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

# Test dobré shody chi-kvadrát

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

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	$X_1$	$X_3$	$X_1+X_3$
<b>slovo B</b>	$X_2$	$X_4$	$X_2+X_4$
<b><math>\Sigma</math></b>	$X_1+X_2$	$X_3+X_4$	$X_1+X_2+X_3+X_4$
	$Np_1$	$Np_3$	
	$Np_2$	$Np_4$	

	žánr C	žánr D	$\Sigma$
<b>slovo A</b>	5	20	25
<b>slovo B</b>	25	20	45
<b><math>\Sigma</math></b>	30	40	70
	10,71	14,29	
	19,29	25,71	

$$\chi^2 = 6,91, \text{ p-hodnota} = 0,009$$

# Test dobré shody chi-kvadrát

- Wikipedia

- [https://cs.wikipedia.org/wiki/Test\\_dobr%C3%A9\\_shody](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: „Test nezávislosti chí-kvadrát by se neměl provádět v případech, kdy ve více než 20 % polí kontingenční tabulky jsou očekávané četnosti menší než 5 a v případech, kdy v některém poli je očekávaná četnost menší než 1.“
- nevhodný pro velká data

	romány	novely	$\Sigma$	% novely
konstrukce A	500000	501800	1001800	50,09%
konstrukce B	501500	500000	1001500	49,93%
$\Sigma$	1001500	1001800	2003300	
chi <sup>2</sup> = 5.43, p=0,020				

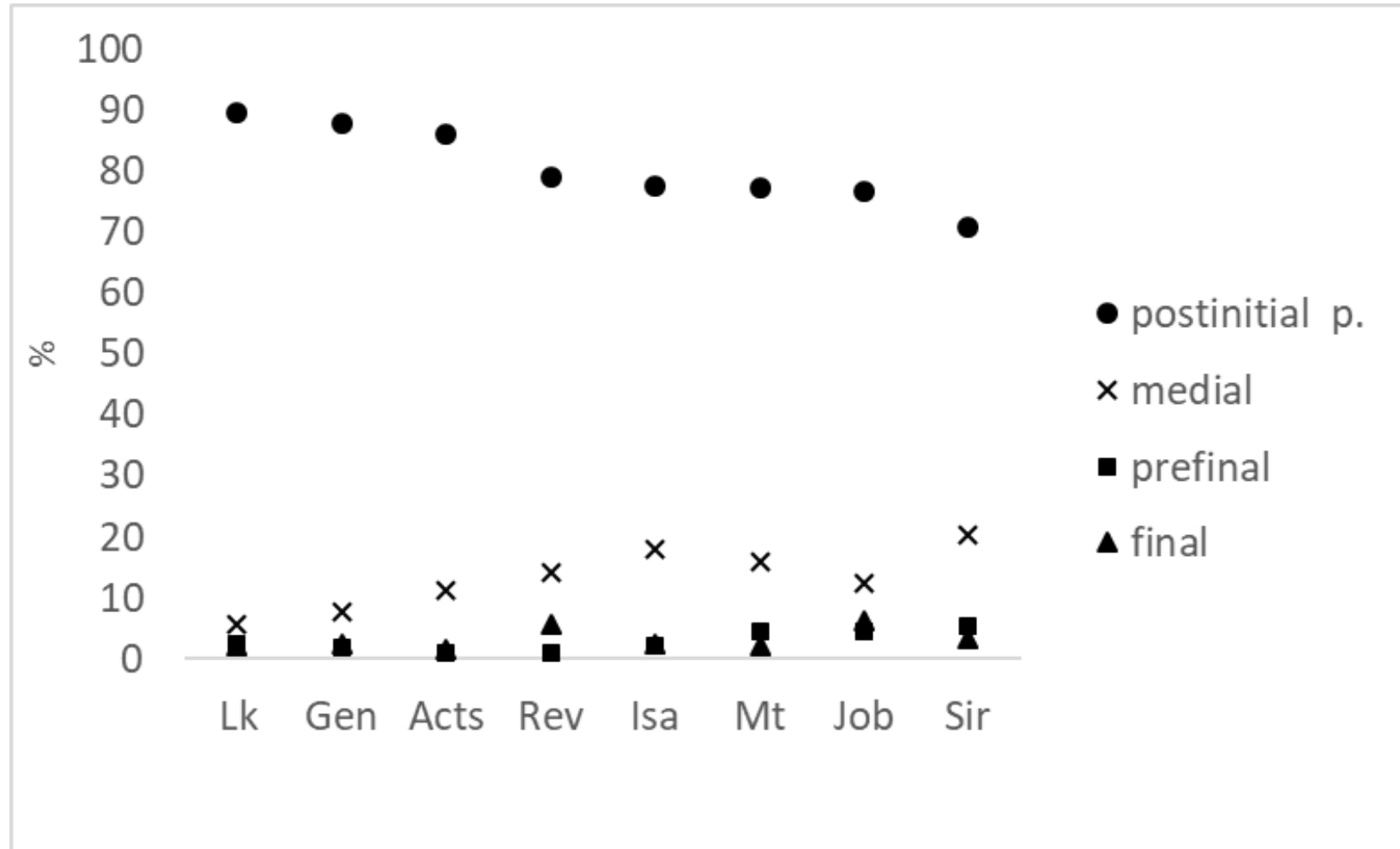
# Příklad: vliv typu textu (žánru) na postavení enklitik

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

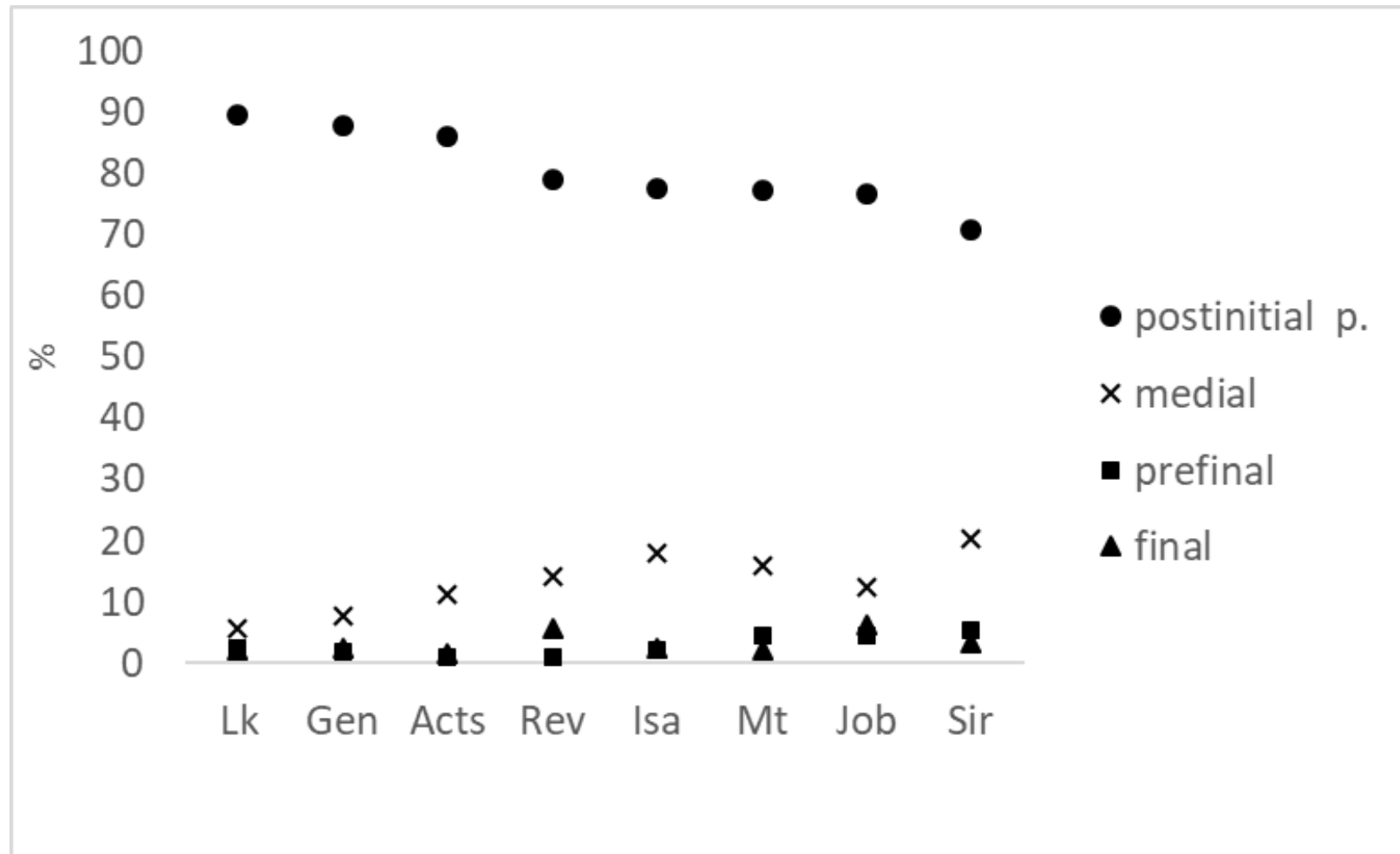
Kosek, P., Navrátilová, O., Čech, R., Mačutek, J. (2018). Word Order of Reflexive 'sě' in Finite Verb Phrases in the First Edition of the Old Czech Bible Translation. (Part 2). *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 135, 3, 189-200.

[http://www.cechradek.cz/publ/2018\\_Kosek\\_etal\\_Krakow\\_j\\_02.pdf](http://www.cechradek.cz/publ/2018_Kosek_etal_Krakow_j_02.pdf)

# Příklad: vliv typu textu (žánru) na postavení enklitik



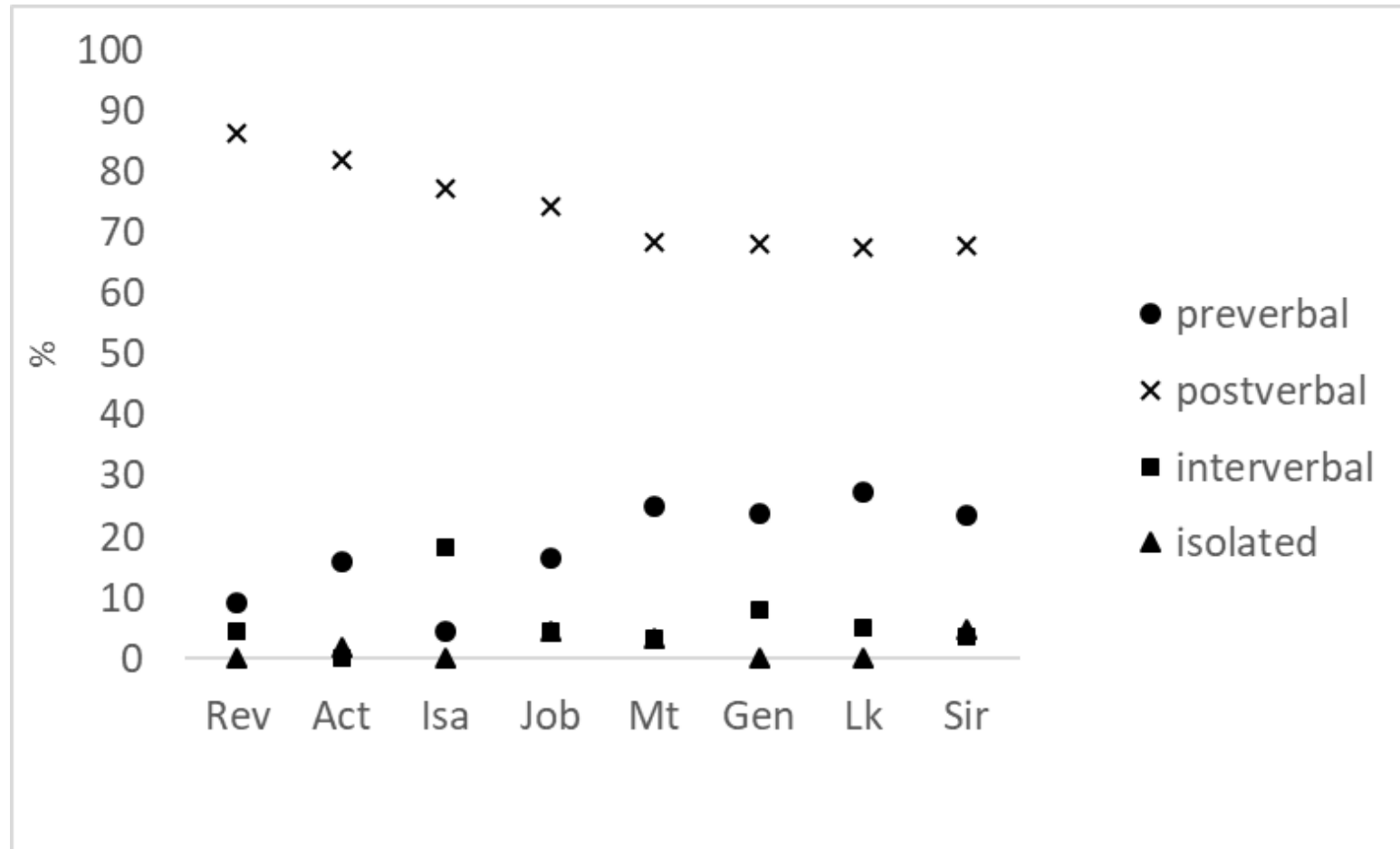
# Příklad: vliv typu textu (žánru) na postavení enklitik



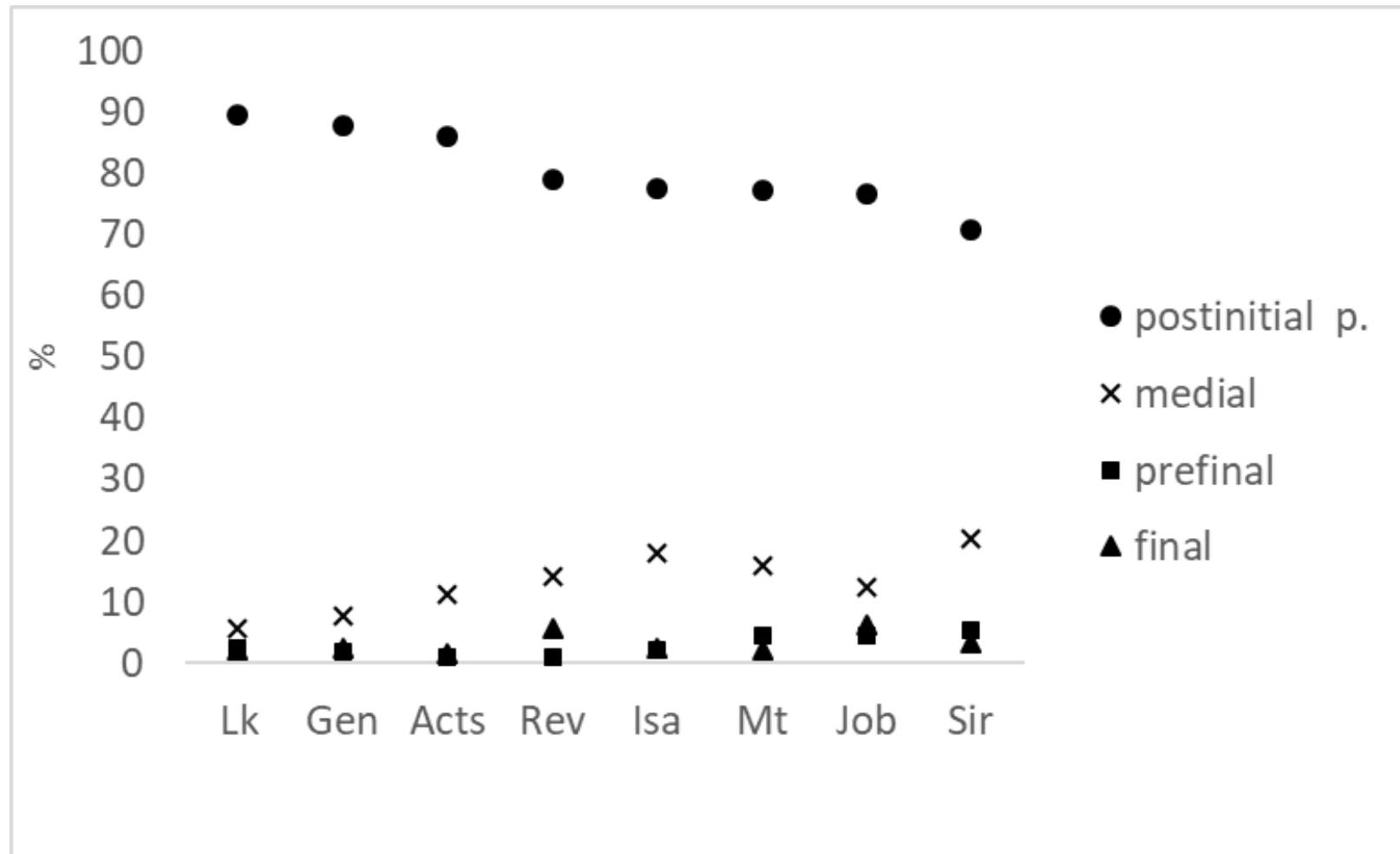
$\chi^2 = 83.712$   
p-value < 0.001



# 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
    - Chi-Square Test Calculator
    - <https://www.socscistatistics.com/tests/chisquare2/default2.aspx>
  - 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?

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

# První pohled?

	<b>děkuji</b>	<b>děkuju</b>
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?

	<b>děkuji</b>	<b>děkuju</b>	<b>% děkuji</b>
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				<i>Row Totals</i>
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
<b>Column Totals</b>	3567	2181				<b>5748 (Grand Total)</b>

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

# Fisherův test

- v případě, že jsou četnosti tak malé, že se nedá chí-kvadrát použít
- <https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickyh-a-biologickyh-dat--biostatistika-pro-matematickou-biologii--testovani-hypotez-o-kvalitativnich-promennych--fisheruv-exaktni-test>