Experimentální obrat v lingvistice

Mojmír Dočekal

2021-11-18

Experimentální lingvistika

2 revoluce

- ve formální lingvistice od druhé poloviny 20. století 2 revoluce:

1. Chomsky (1957): kognitivní a formální přístup k syntaxi:

the whole goal of science is to replace complex visibles by simple invisibles. That is science. If you are not doing that, then it is something else, it is data organization, flower collection. Sometimes the latter is useful, but it should not be confused with science. If it is science, since Galileo, it is an effort to satisfy Galileo's maxim: nature is simple. If we have not figured it out, it is our problem.

2. 90.léta: experimentální obrat v lingvistice (klasické učebnice: Baayen (2008), Kruschke (2011))

- možná ještě důležitější
- analýza velkého množství dat (ať už uložených: korpusy, nebo experimentálně získaných)
- dost často jde o subtilnější a více variovaná data než v syntaxi dříve:
- syntaktický a jasný rozdíl:

(1) a. Já jsem nepřišel.
b. *Já nejsem přišel.

- názorný příklad (z NESČ):
- neutralizace časových a aspektuálních rozdílů v negovaných větách
- příklad z jazyka *bafut* (Kamerun)
- mizí rozdíl mezi přítomným perfektem a nedávnou minulostí v negovaných větách

(2) a. mbìŋ lòó
déšť padal
‘Pršelo/Měli jsme deštivo.’

b. mbìŋ lòò me’
déšť padal IMPST
‘Pršelo/Právě pršelo.’

c. kāā mbìŋ sì lōò
NEG déšť NEG padal
‘Nepršelo.’

- neutralizace aspektového rozdílu: pod negací musíme použít default (časový nebo aspektový)

- podobně pro češtinu se tvrdí, že (Hajičová z NESČ):

O souhře č. slovesné n. a vidu se zpravidla říká, že dok. sloveso v imper. má ve své základní funkci (zákaz, záporná rada aj.) jako přímý záporný protějšek sloveso nedok.: *Sedni si dopředu! – Nesedej si dopředu!* Jde však o jev širší, protože obdobně se chová i sloveso durativní ve vztahu k iterativnímu: *Jeďte zítra do Pardubic! – Nejezděte zítra do Pardubic!* Kromě toho nejde jen o imper., ale i o různé významově příbuzné vazby: *Měli byste jet do Pardubic – Neměli byste jezdit do Pardubic;Rád by jel do Pardubic – Nerad by jezdil do Pardubic*. Hranice tohoto jevu dosud nebyly s plnou soustavností prostudovány (viz přehled popisů v Karlík and Nübler (1998)), podobně jako není jasné, do jaké míry platí o užití imper. ve smyslu varování (výstrahy), že takové variaci nepodléhá. Jistě to platí např. o *Nesedni si na klobouk!*, ale není to docela jasné u ostatních příkladů n. u vět jim podobných.

- pokud chceme popsat takové jevy, tak musíme pracovat nejen s intuicí
- a používat nástroje pro práci s nejednoznačnými daty (signál a šum: https://en.wikipedia.org/wiki/The_Signal_and_the_Noise)
- příklad z historie: Galton a jeho studie o dědičnosti výšky
- podobný vzor: intuitivně jasná korelace, ale spousta protipříkladů

Francis Galton

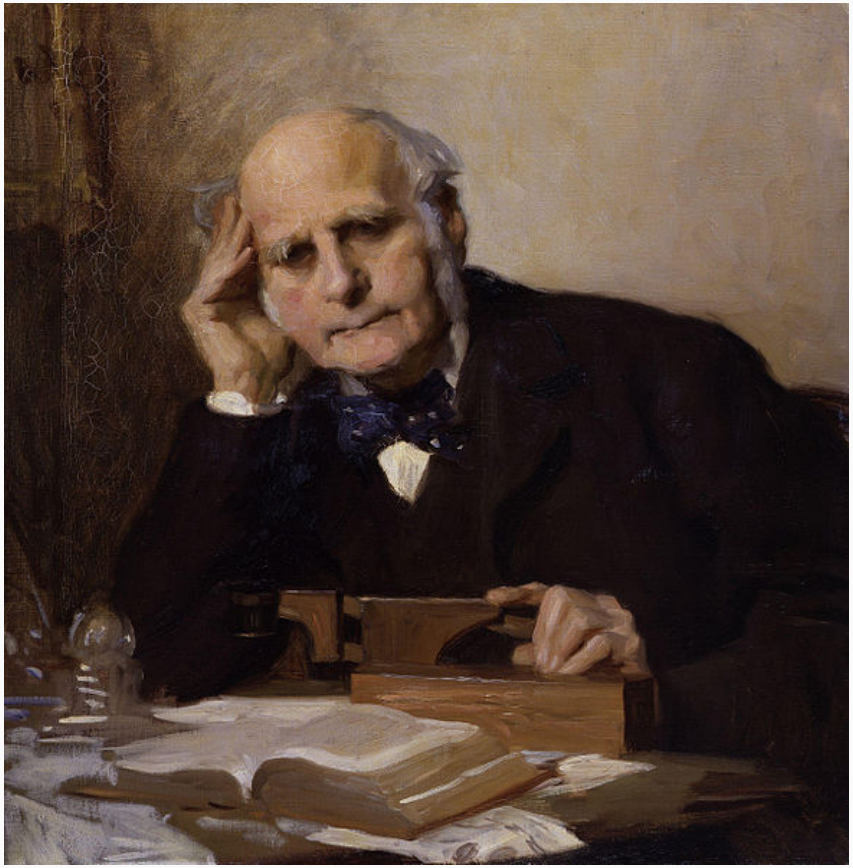


Figure 1: Galton

- cestovatel, antropolog, eugenik
    - základy deskriptivní statistiky (median – vox populi) a inferenční statistiky (lineární regrese)
    - také meteorolog, statistik (efektivita modliteb) a poměřovatel krásy žen v různých částech Anglie
- viktoriánský učenec s vášní pro data
- bratranec Charlese Darwina
- výzkumná otázka:

(3) Známe-li výšku rodičů, lze předpovědět výšku jejich dětí?

- výzkumná otázka se ve statistice vždy staví proti nulové hypotéze:

(4) Mezi výškou rodičů a výškou dětí není žádný vztah.

- Galton: shromáždil data o cca 400 rodičích a jejich 400 dcerách a synech

- databáze
- 1 palec = 2.54 cm
- dnešní průměrná výška mužů a žen v UK je 69 a 63 palců (vs. 69.5 a 64 medián v Galtonově vzorku)
- dál statistické zpracování z R Core Team (2021)
- napřed deskriptivní statistika
- následuje výstup z jazyka R (v české mat. terminologie se používá desetinná čárka)

```
## [1] "Fathers: počet záznamů, deskriptivní stat."

## [1] "standard deviation (směrodatná odchylka)"

## [1] 197

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   62.00   68.00   69.50   69.35   71.00   78.50 

## [1] 2.622034

## [1] "Mothers"

## [1] 197

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   58.00   62.70   64.00   63.98   65.50   70.50 

## [1] 2.355607
```

```
## [1] "Sons"
```

```
## [1] 465
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   60.00   67.50   69.20   69.23   71.00   79.00
```

```
## [1] 2.631594
```

```
## [1] "Daughters"
```

```
## [1] 433
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.00   62.50   64.00   64.11   65.50   70.50
```

```
## [1] 2.37032
```

- quantile: česky dolní kvantil, medián, horní kvantil

- důležité termíny: **mean** (průměr) vs. **median** (medián)
- směrodatná odchylka

```
x <- c(2,4,8,10,100)
mean(x)
```

```
## [1] 24.8
```

```
median(x)
```

```
## [1] 8
```

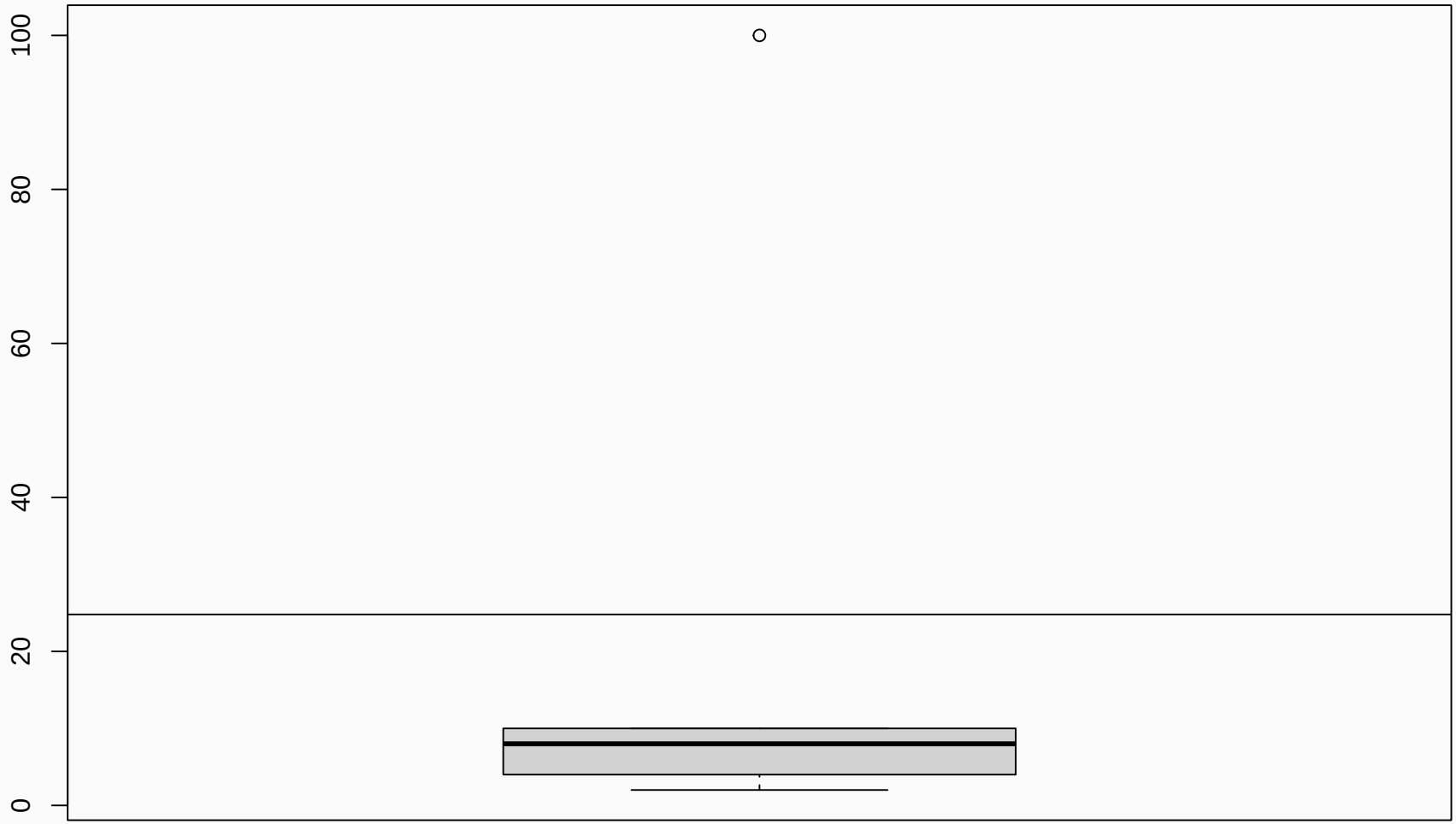
```
sd(x)
```

```
## [1] 42.15685
```

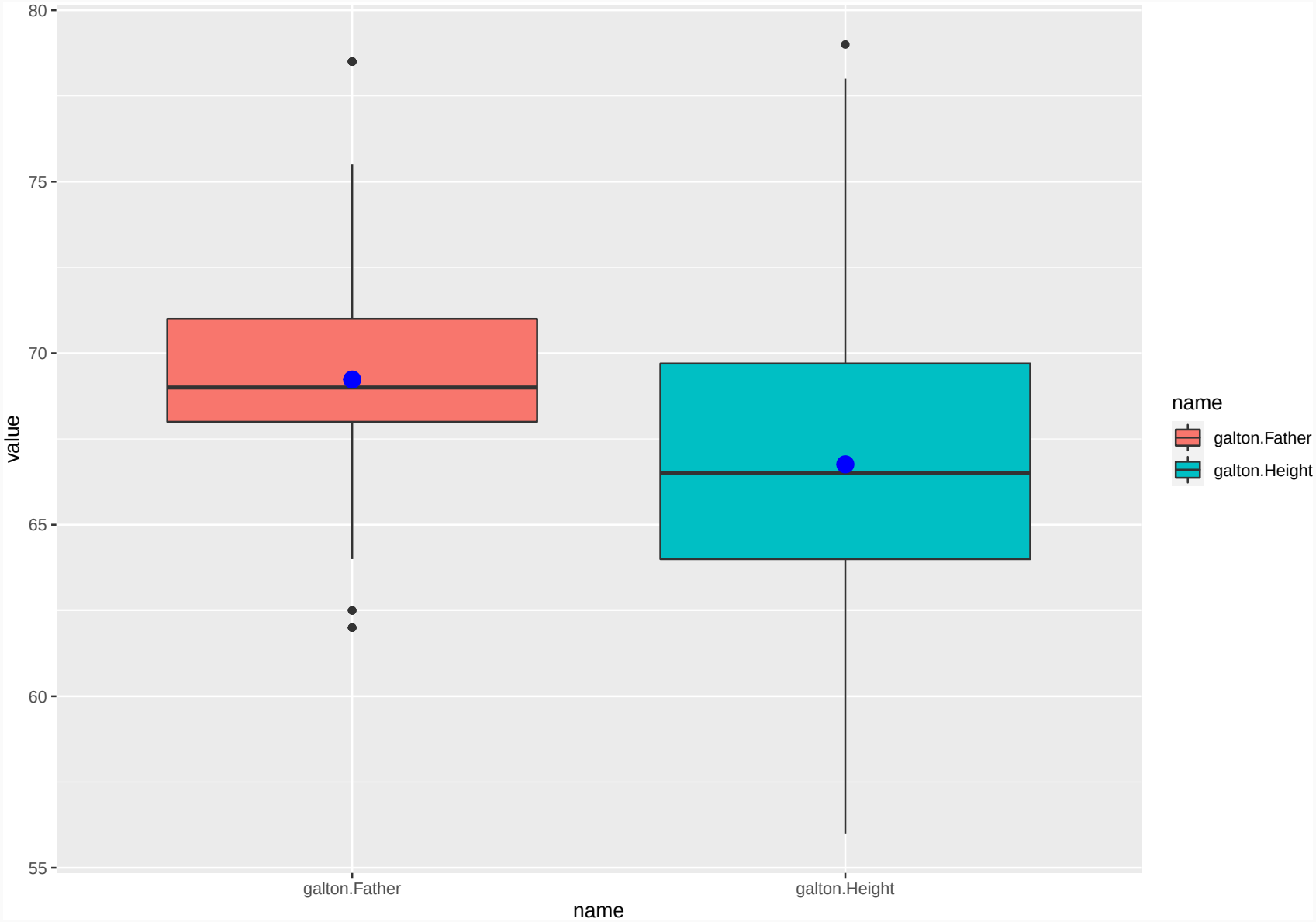
```
y <- c(2,4,8,10)
sd(y)
```

```
## [1] 3.651484
```

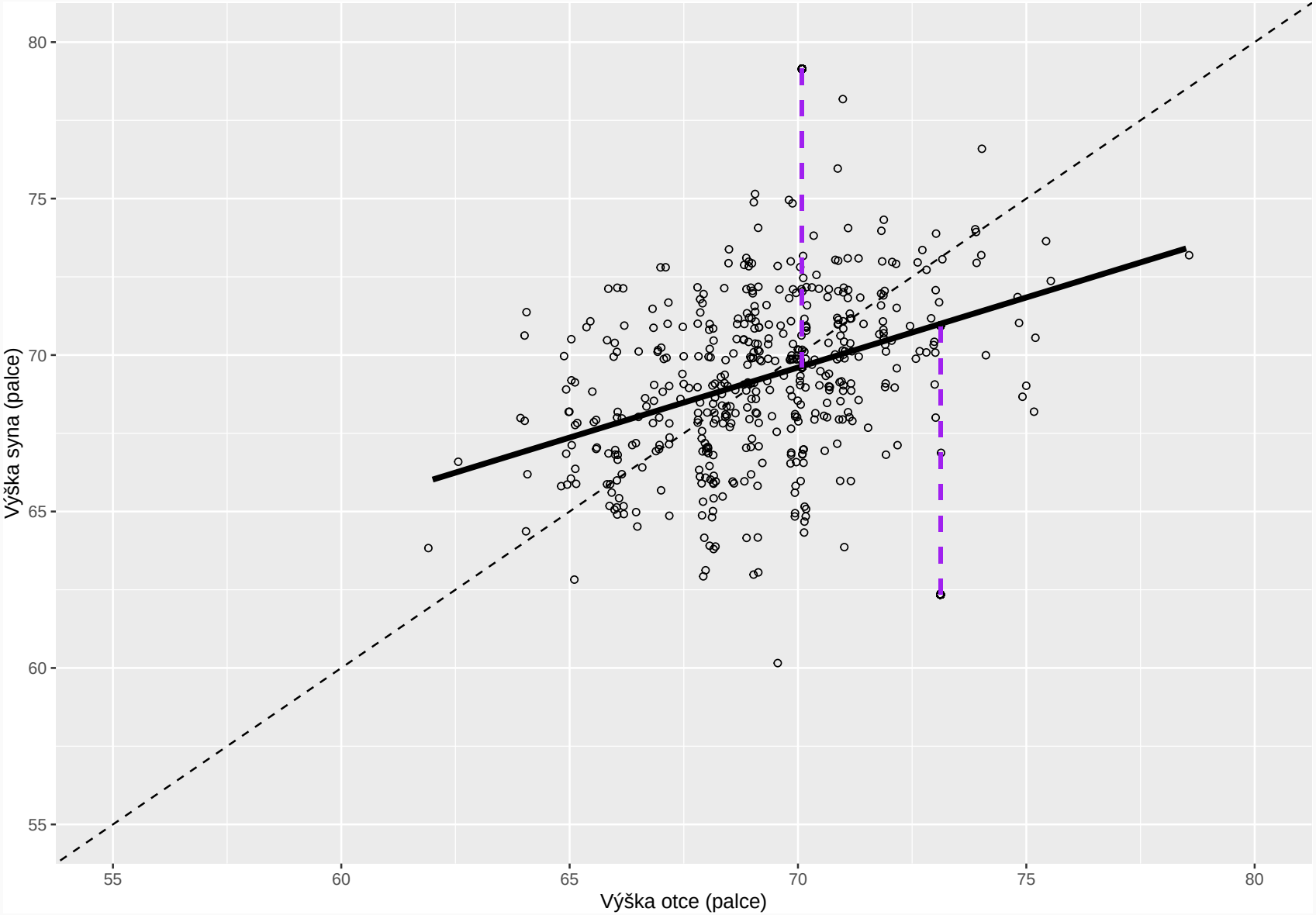
```
## [1] "Boxplot (krabicový graf)"
```



- obvyklé grafické znázornění: tzv. boxplot nebo scatter plot (korelační diagram)
- deskriptivní statistika


80
75
70
65
60
55
value
galton.Father
galton.Height
name
name
galton.Father
galton.Height

- následující slide: tzv. scatter-plot graf výšky otců (x) oproti výšce synů (y)
- přidaný jitter pro odlišení stejných hodnot
- přerušovaná čára: výška otce = výška syna
- tlustá čára (regresní přímka): lineární regrese, tzv. best fit (nejlepší aproximace?)
- reziduální chyba: vzdálenost bodu od lineárně regresní přímky
- podle Spiegelhalter (2019)


Výška syna (palce)
80
75
70
65
60
55
55
60
65
70
75
80
Výška otce (palce)

- moderní interpretace pomocí tzv. lineárního modelu:
- Estimate (odhad): jak se změní výška syna, vzroste-li explanatory (vysvětlující?) proměnná (výška otce) o 1 (palec), plus intercept (průsečík)
- t-value: stejné jako Studentův t-test: jak daleko je estimate od 0 měřeno standardními chybami (více než abs. 2 – statisticky signifikantní (významný) efekt)
- p-hodnota: pravděpodobnost nulové hypotézy

```
##
## Call:
## lm(formula = Son ~ FatherS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3774 -1.4968  0.0181  1.6375  9.3987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.25891    3.38663   11.30   <2e-16 ***
## FatherS      0.44775    0.04894    9.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.424 on 463 degrees of freedom
## Multiple R-squared:  0.1531, Adjusted R-squared:  0.1513
## F-statistic: 83.72 on 1 and 463 DF,  p-value: < 2.2e-16
```

- to už je inferenční (úsudková) statistika (induktivní úsudek ze vzorku na populaci)
- přesně tento typ modelů (ale smíšených) pak využíváme v lingvistice: Mariia Onoeva: Inferential behavior of degree achievements: an experimental study

- ilustrace inferenční statistiky
- https://www.britannica.com/topic/regression-to-the-mean
- seriózní analýza by musela vzít v úvahu česká data: http://www.szu.cz/uploads/documents/obi/CAV/6.CAV_2_Dlouhodobe_zmeny_rustu.pdf

```
MDI <- 194/2.54
over_median <- MDI - 69.5
over_median
```

```
## [1] 6.877953
```

```
MDII <- 69.2 + over_median*0.45
MDII
```

```
## [1] 72.29508
```

```
metric <- MDII*2.54
metric
```

```
## [1] 183.6295
```

Zpět k lingvistice

- příklad s neutralizací aspektu v negovaných imperativních větách
- výzkumná hypotéza v (5)
- oproti tomu nulová hypotéza v (6):

(5) V českých imperativech dochází u negovaných sloves k neutralizaci vidového rozdílu (sloveso je použito v defaultním, tj. imperfektivním vidu).

(6) Negace nemá vliv na neutralizaci vidového rozdílu.

- první krok: data

- pro tento typ dat je vhodný Český národní korpus
- hledání slov, frází:

1. kernel

2. širší než delší

- hledání podle morfologických značek (tag):

1. baseline (základní případ?):

a. počet imperativních imperfektiv: `[tag="Vi.............I"]`: 104 385

b. počet imperativních perfektiv: `[tag="Vi.............P"]`: 165 924

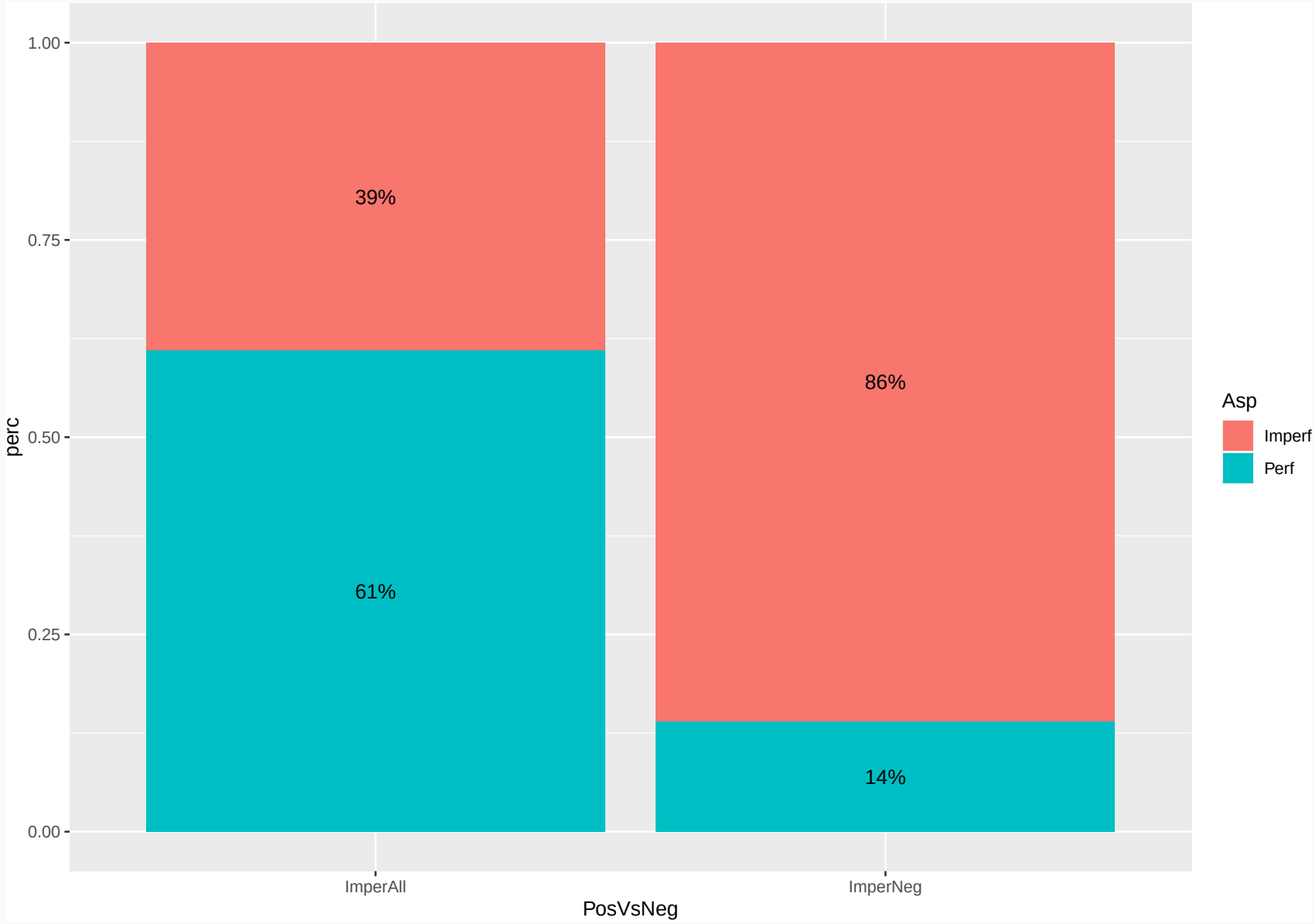
(oproti zbytku převažují perfektiva: 11 514 033 (Imperf.) vs. 5 787 695 (Perf.))

2. počty negovaných perfektiv a imperfektiv:

a. negovaná imperfektiva: `[tag="Vi.........N....I"]`: 27 256

b. negovaná perfektiva: `[tag="Vi.........N....P"]`: 4 567

- grafy: boxplot relativní frekvence (krabicový graf relativní četnosti)


1.00
0.75
0.50
0.25
0.00
perc
39%
61%
86%
14%
ImperAll
ImperNeg
PosVsNeg
Asp
Imperf
Perf

- k inferenční statistice (od vzorku k populaci):
- u těchto tzv. count (spočetných?) dat je nejobvyklejší způsob testování přes Fisherův nebo chi square test (rozdělění chí-kvadrát)
- Fisherův test:

```r
data <- matrix(c(27256, 104385, 4567, 165924), ncol = 2, dimnames = list(c("ImperAl
fisher.test(data)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  data
## p-value < 2.2e-16
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  9.182734 9.796871
## sample estimates:
## odds ratio
##   9.487635
```

- pravděpodobnost, že s takovými daty je kompatibilní nulová hypotéza: *p-value < 2.2e-16*
- a pro negovaný imperativ je 9.49 krát pravděpodobnější, že sloveso bude v imperfektivu (oproti všem imperativům)
- confidence interval (interval spolehlivosti): 95% pravděpodobnost, že jakýkoliv jiný náhodný vzorek v populaci se bude chovat stejně: 9.182734 – 9.796871
- tj. pravděpodobnost *nepřečti* v češtině oproti *nečti* leží někde v tomto 95% confidence intervalu

Příklad minimálního experimentu

- při překřížení rukou někteří lidí: pravá nahoře, ...
- výzkumná otázka: má biologický rod vliv na to která?

(7) Nulová hypotéza: mezi oběma proměnnými není žádný vztah.

- kód: RstudioCloud

References i

Baayen, H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Chomsky, Noam. 1957. *Syntactic Structures*. the Hague: Mouton.

Kruschke, John K. 2011. *Doing Bayesian Data Analysis : A Tutorial with r and BUGS*. Burlington, MA: Academic Press. http://www.amazon.com/Doing-Bayesian-Data-Analysis-Tutorial/dp/0123814855.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spiegelhalter, David. 2019. *The Art of Statistics: Learning from Data*. London: Penguin books.