Form. a exper. sémantika II/Exper. syntax a sémantika II

JS 2023

Mojmír Dočekal & Lucia Vlášková
ÚJABL MUNI

24/02/2023

Cíle, ukončení, literatura, obsah

Intro

- vzájemné představení, …
- materiály: v ISu

Cíle

- kurz o statistice, formální lingvistice a experimentech
- dílčí cíle:
  - příprava k samostatné práci na lingvistickém experimentu (diplomka)
  - lingvistika a data science
  - soft-skills: statistika, grafy, R (R Core Team (**2019**)), …

Úvod

Proč statistika v lingvistice?

- čím jsou data víc plná šumu, tím víc je potřeba statistika
- v lingvistice:

a. Petr neviděl nikoho.

b. *Petr viděl nikoho.

c. ne-roz-šiř-ova-t

d. …

- oproti:

a. Petr nechtěl, aby nikdo přišel.

b. Petr nechtěl, aby přišel ani jeden student.

c. Aleš a Bára věří, že na půdě FF žijí dva duchové. (kumulativní čtení?)

Statistika

- statistika je kontra-ituitivní
- lidé mají dobré intuice o:
  - gramatičnosti
  - vyplývání
  - aritmetice
  - nikdy o statistice (Kahneman & Tversky, Koralus & Mascarenhas)

Ilustrativní příklad:

- Bayesovská statistika
- **Thomas Bayes**
- **Signál a šum, kap. 8**
- náhled do RStudia

Lehké připomenutí statistiky

- pravděpodobnost: frekventistická verze
- pravděpodobnost narození syna nebo dcery
- součet: 1
- vtipy: meteorolog předpovídající 50% pravděpodobnost deště na sobotu i na neděli: 100% deště během víkendu, muž s 2 bombami v letadle

Konjunkce, disjunkce, negace a podmíněná pravděpodobnost

- pravděpodobnostní protějšky logických spojek

Konjunkce

- mají-li Novákovi dvě děti, jaká je pravděpodobnost toho, že jsou obě dcery?
- obecně: $p(A \wedge B) = p(A) * p(B)$

- obě události muzí být nezávislé
- ilustrace problému: **Meadowův zákon**

Disjunkce pravděpodobností

- mají-li Novákovi 2 děti, jaká je pravděpodobnost, že alespoň jedna z nich je dívka?
- obecně: $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$
- případně přes počítání kombinací
- počasí: 1/2 + 1/2 - (1/2 * 1/2)
- pravděpodobnosti vzájemně exkluzivních událostí se sčítají: v sobotu buď prší nebo neprší, …

Pravděpodobnost komplementu/negace události

- 1-p(A)
- pravděpodobnost toho, že Novákovi mají alespoň 1 dceru = (1 - p(dva synové)) … (1-0.25)
- pravděpodobnost toho, že za 10 let nevypukne válka (baserate: 10% každý rok) = (1 - 0.9 x 0.9 … 0.9) = (1 - 0.35)

Podmíněná pravděpodobnost

- nejsložitější

  xkcd

- p(zabit-v-bouřce|mimo-bouřku) $\neq$ p(zabit-v-bouřce|v-bouřce)

- příklad: je-li Novákovic starší dítě dívka, jaká je pravděpodobnost, že mají dvě dcery?
- obecně: $p(a|b) = p(a \wedge b)/p(b)$

- ještě komplikovanější: Bayesova věta (založená na podmíněné pravděpodobnosti)

```
library(LaplacesDemon)

help(BayesTheorem)

# Pr(Hell|Consort) =
PrA <- c(0.75,0.25)
PrBA <- c(6/9, 5/7)
(BayesTheorem(PrA, PrBA))
```

```
[1] 0.7368421 0.2631579
attr(,"class")
[1] "bayestheorem"
```

```
# > [1] 0.7368421
```

- Bayes teorem aplikovaný na klasický případ chybného úsudku podmíněné pravděpodobnosti:

1. 1% žen v populaci má rakovinu prsu
2. citlivost testu (true-positive) je 90%
3. false-positive je 9%
4. Má-li žena pozitivní test, jaká je pravděpodobnost toho, že má opravdu rakovinu?

   (nejpopulárnější odpověď lékařů: 80-90%)

```r
library(LaplacesDemon)

help(BayesTheorem)

# Pr(Nemoc|Test) =
PrA <- c(0.01,0.99)
PrBA <- c(9/10, 0.9/10)

BT <- BayesTheorem(PrA, PrBA)

print(paste("V procentech (pravděpodobnost nemoci): ", round(B
```

```
[1] "V procentech (pravděpodobnost nemoci):  9 %"
```

- a obráceně:

```
# Pr(Test|Nemoc) =
PrA <- c(9/10, 0.9/10)
PrBA <- c(0.01,0.99)

# (0.9*0.99)/0.99

BT <- BayesTheorem(PrA, PrBA)
print(paste("V procentech (pravděpodobnost nemoci): ",round(BT
```

```
[1] "V procentech (pravděpodobnost nemoci):  91 %"
```

Dokumentace k Bayesově větě (LaplacesDemon)

- RDocumentation

Nejznámější chyby ve statistických inferencích

Konjunkce

(připomenutí)

- mají-li Novákovi dvě děti, jaká je pravděpodobnost toho, že jsou obě dcery?
- obecně: $p(A \wedge B) = p(A) * p(B)$

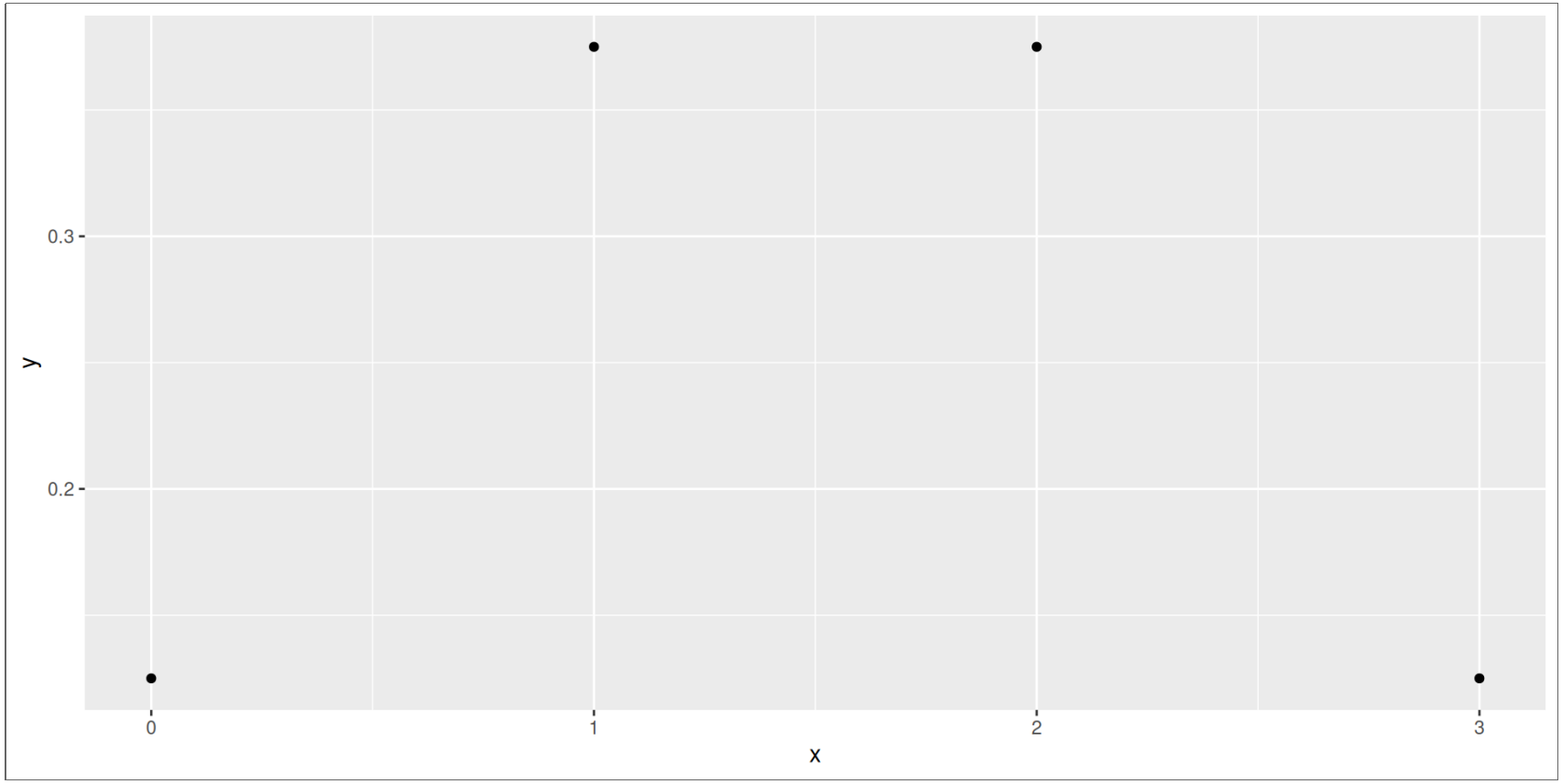
- obecně: pro *x* pokusů z *n* je pravděpodobnost určena binomickým rozdělením
  - typ probability mass function (pravděpodobnostní funkce) pro diskrétní proměnné

```r
x <- seq(0,3)

y <- dbinom(x, 3, prob = 1/2)

df <- data.frame(x,y)

df
```

```
  x     y
1 0 0.125
2 1 0.375
3 2 0.375
4 3 0.125
```

- např. pravděpodobnost toho, že se narodí dvě dcery v rodině s třemi dětmi

```
1 library(ggplot2)
2
3 qplot(x,y)
```



- pomocí frekvenčního stromu: narození jedné dcery je možné 3 způsoby, tj. stejná pravděpodonbost (3/8=0.375)
- ale pravděpodobnost narození jedné dcery je větší než narození dvou dcer (pro rodinu s 2 dětmi)

```r
x <- seq(0,2)

y <- dbinom(x, 2, prob = 1/2)

df <- data.frame(x,y)

df
```

```
  x    y
1 0 0.25
2 1 0.50
3 2 0.25
```

- obecně:

  $p(A \wedge B) \leq p(A)$

- z Kahneman (**2011**)
- kontext:

  Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

- úkol subjektu: následujících 8 scénářů seřadit podle pravděpodobnosti (nebo reprezentativnosti):

1. Linda is a teacher in elementary school.
2. Linda works in a bookstore and takes yoga classes.
3. Linda is active in the feminist movement.
4. Linda is a psychiatric social worker.

5. Linda is a member of the League of Women Voters.
6. Linda is a bank teller.
7. Linda is an insurance salesperson.
8. Linda is a bank teller and is active in the feminist movement.

- nejdůležitější kontrast:

6. Linda is a bank teller.

   vs.

8. Linda is a bank teller and is active in the feminist movement.

- původním designu byl kontrast testován between-subjects: 7 podmínek, jen 1 z 6 xor 8
- výsledek: všichni testovaní seřadili 8 s větší pravděpodobností než 6
- nový experiment: within-subject (v podstatě jen kontrola) – tj. všech 8 podmínek
- výsledek znovu 8 více pravděpodobné než 6

- chyba statistického uvažování: $p(A \wedge B) \not> p(A)$ ($p(A \wedge B) \not> p(B)$)
- nebo jasněji: $p(A \wedge B) \leq p(A)$
- Kahnemanovo vysvětlení: chyba Systému 2, Systém 1 je natolik silný, že zablokuje Systém 2
- výsledky (80. leta): 89% BA studentů chyba, stejné pro Stanford Graduate School of Business – 85% chybovost

- pokusy o nápravu: redukce na pouhé dvě podmínky:

1. Linda is a bank teller.

   vs.

2. Linda is a bank teller and is active in the feminist movement.

- úsudková chyba stále zůstala u 85% - 90% BA studentů
- Stephen Jay Gould: He knew the correct answer, of course, and yet, he wrote, “a little homunculus in my head continues to jump up and down, shouting at me—‘but she can’t just be a bank teller; read the description.’”

- možná cesta (psychologické řešení) ven: převedení otázky na relativní frekvenci:

  Imagine a thousand women like Linda. How many of them do you think are bank tellers? How many of them do you think are bank tellers who are active in the feminist movement?

- lepší výsledky: Tversky and Kahneman (**1983**),Hertwig and Gigerenzer (**1999**)
- lingvistické řešení: Asudeh and Giorgolo (**2020**)

Méně známý problém: chyby v usuzování o podmíněné pravděpodobnosti

- z Kahneman et al. (**1982**)
- scénář: svědek nehody způsobené taxíkem pozdě v noci
- dvě taxikářské společnosti: Zelené taxi (85%), Modré taxi (15%) – base rate (priors)
- svědek prohlašuje, že taxi bylo modré

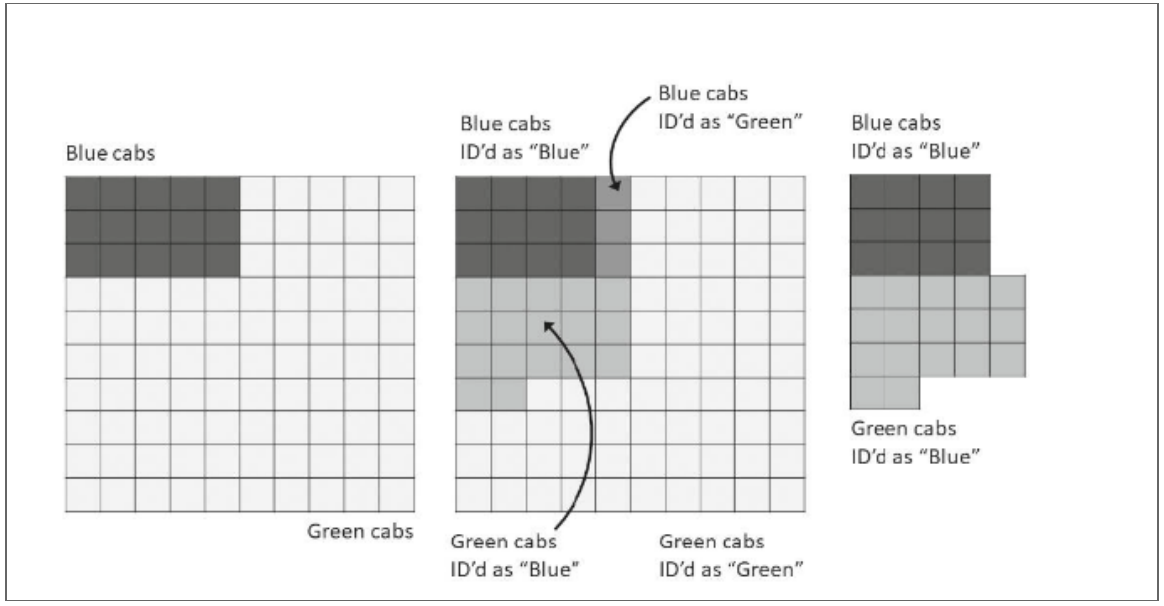
- svědek na testu prokázal 80% spolehlivost k rozeznání barvy za stejných podmínek
- jaká je pravděpdobnost p(Modré|priors)

- base rate (priors): pravděpodobnost hypotézy
- p(Data|Hypotéza): likelihood – jak pravděpodobná by byla data, pokud by hypotéza byla pravdivá
  - tj. ze 100 náhodných taxíků (85 zelených, 15 modrých) by svědek identifikoval 20 jako modrých (20% chybovost)

```
1  library(LaplacesDemon)
2
3  # Pr(Hypothesis|Data) =
4  PrHypothesis <- c(0.85,0.15)
5  PrDataHypothesis <- c(2/10, 8/10)
6
7  HypothesisData <- BayesTheorem(PrHypothesis, PrDataHypothesis)
8
9  print(paste("V procentech (pravděpodobnost modrého taxíku): ",
```

```
[1] "V procentech (pravděpodobnost modrého taxíku):  41 %"
```

- v původním experimentu byla odpověď (median): 80% ve prospěch hypotézy (byl to modrý taxík)
- tj. dvakrát víc než správná odpověď
- psychologická odpověď: převést na frekvence a ještě ulehčit vizualizací



Pinker (**2022**)

Příklady lingvistických studií užívajících:

1. frekventistická analýza projekce presupozic: **Chemla 2009**

- Chemla (**2009**)

2. bayesovská statistika skalárních implikatur: **Franke & Jäger 2016**

References

Asudeh, Ash, and Gianluca Giorgolo. 2020. *Enriched Meanings: Natural Language Semantics with Category Theory*. Vol. 13. Oxford University Press.

Chemla, Emmanuel. 2009. “Presuppositions of Quantified Sentences: Experimental Data.” *Natural Language Semantics* 17 (4): 299–340.

Hertwig, Ralph, and Gerd Gigerenzer. 1999. “The ‘Conjunction Fallacy’revisited: How Intelligent Inferences Look Like Reasoning Errors.” *Journal of Behavioral Decision Making* 12 (4): 275–305.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. macmillan.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge university press.

Pinker, Steven. 2022. *Rationality: What It Is, Why It Seems Scarce, Why It Matters*. Penguin.

R Core Team. 2019. *R: A Language and Environment for Statistical*

Computing. Vienna, Austria: R Foundation for Statistical Computing. **https://www.R-project.org**.

Tversky, Amos, and Daniel Kahneman. 1983. “Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment.” *Psychological Review* 90 (4): 293.

Error ×