Rychlý průlet Bayesovskou statistikou

Aleš Vomáčka

Dnešní cíle

1) Rozdíly mezi frekventistickou a bayesiánskou statistikou
2) Jak bayesiánská statistika funguje
3) Jak na baysiánskou statistiku v R

Paradigmata ve statistice

Paradigmata ve statistice

- Statistika má paradigmata stejně jako sociologie
- Dnes je nejpopulárnější frekventistická statistika
  - … která nám dala p hodnoty, testování nulových hypotéz apod.
- Bayesiánská statistika ale nabírá na síle (a navíc byla první)

Paradigmata ve statistice

- Dva základní rozdíly mezi frekventisty a bayesiánci:
  - Co je pravděpodobnost?
  - Jak se dívat na populační parametry?

Co je pravděpodobnost?

Co je pravděpodobnost?

- **Frekventisti:** Pravděpodobnost je (limitní) výskyt jevu napříč velkým počtem pokusů.
- **Bayesiánci:** Pravděpodobnost je subjektivní míra jistoty, že dojde k nějakému jevu (kterou můžeme interpretovat jako sázkařský kurz).

Co je pravděpodobnost?

- Jak interpretovat $P(mince = orel) = 0.9$?
- **Frekventisti:** Pokud bychom mnohokrát hodili touto mincí, v 9 případech z 10 by padnul orel.
- **Bayesiánci:** Jsme si dost jistý, že pokud bych hodil touhle mincí, tak by padnul orel, abych si na to vsadil s kurzem 9:1.

Co je pravděpodobnost?

- Frekventistická definice pravděpodobnosti objektivní, zaměřená na opakování pokusů
- Bayesiánská definice pravděpodobnosti subjektiní (ale ne arbitrátní), zaměřená na kvantifikaci osobních znalostí

Jak se dívat na populační parametry?

Jak se dívat na populační parametry?

- Populační parametr = skutečná hodnota, kterou hledáme
  - (např. volební prefrence dané strany v populaci)

- **Frekventisti:** Populační parametr je jedna fixní (většinou neznámá) hodnota.

- **Bayesiánci:** Populační parametr je hodnota, která se s určitou pravděpodobností vyskytuje v daném intervalu.

Jak se dívat na populační parametry?

- Interpretace populačních parametrů (a pravděpodobností) v praxi důležitá pro interpretaci intervalových odhadů.

- **Frekventisti:** Interval spolehlivost je interval, který bude napříč velkým počtem vzorků obsahovat skutečnou hodnotu (populační parametr) s danou pravděpodobností.

- **Bayesiánci:** Interval kredibility je interval, který s danou pravděpodobností obsahuje skutečnou hodnotu (populační parametr).

Jak se dívat na populační parametry?

- Jak interpretovat, že podpora strany je 95% CI $[26; 29]$ procent?

- **Frekventisti:** Pokud bychom opakovaně tahali velké množství vzorků stejné velikosti z populace a pro každý z nich spočítali 95% interval spolehlivosti, 95 % z nich by obsahovalo skutečnou hodnotu podpory strany. Rozpětí těchto intervalů by mělo cca 3 procentní body (29 - 26).

- **Bayesánci:** Na 95 % je skutečná podpora strany něco mezi 26 a 29 procenty.

Shrnutí

- **Frekventistická statistika** založená myšlence opakovaného měření, garantuje určité vlastnosti napříč velkým množstvím vzorků (např. pokrytí intervalů spolehlivosti), ale neříká nic o výsledcích konkrétních experimentů nebo našich hypotézach

- **Bayesiánská statistika** založená na kvantifikaci našich představ o světě, umožňuje přiřazovat pravděpodobnost našim hypotézám, ale negarantuje nic napříč velkým množsvím experimentům.

O čem je bayesiánská statistika

Bayesův teorém

- Centrem bayesiánské statistiky je bayesův teorém:

$$P(Hypotza|Data) = \frac{P(Data|Hypotza) * P(Hypotza)}{P(Data)}$$

- V praxi se dá zjednodušit:

$$P(Hypotza|Data) \propto P(Data|Hypotza) * P(Hypotza)$$

Bayesův teorém

$$P(Hypotza|Data) \propto P(Data|Hypotza) * P(Hypotza)$$

- $P(Hypotza)$ je **prior**, naše dosavadní představa o světě.
- $P(Data|Hypotza)$ je **likelihood**, informace obsažené v datech, které máme k dispozici.
- $P(Hypotza|Data)$ je **posterior**, naše představa o světě obohacná o data, která máme k dispozici.

Bayesův teorém

- Všimněte si, že posterior je kombinace (kompromis) toho, co už jsme věděli a nových důkazů, které máme k dispozici.

$$Posterior \propto Likelihood * Prior$$

- To nás vede k (podle mě) největší taháku bayesiánské statistiky...

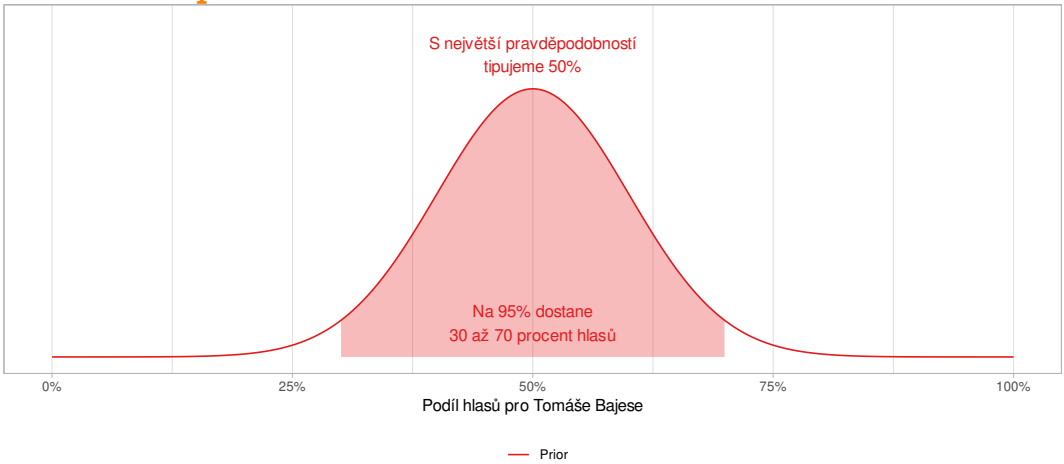
Bayesiánská statistika je formalizovaný způsob, jak aktualizovat naše názory na svět.

Příklad s volbami

Příklad - Jak aktualizovat náš pohled na svět

- Představte si, že je rok 2023 a blíží se prezidentské volby.
- Slyšeli jste, že se objevil nový kandidát jménem Tomáš Bajes.
- Je prý docela populární.
- Kolik procent voličů ho podle vás bude volit?

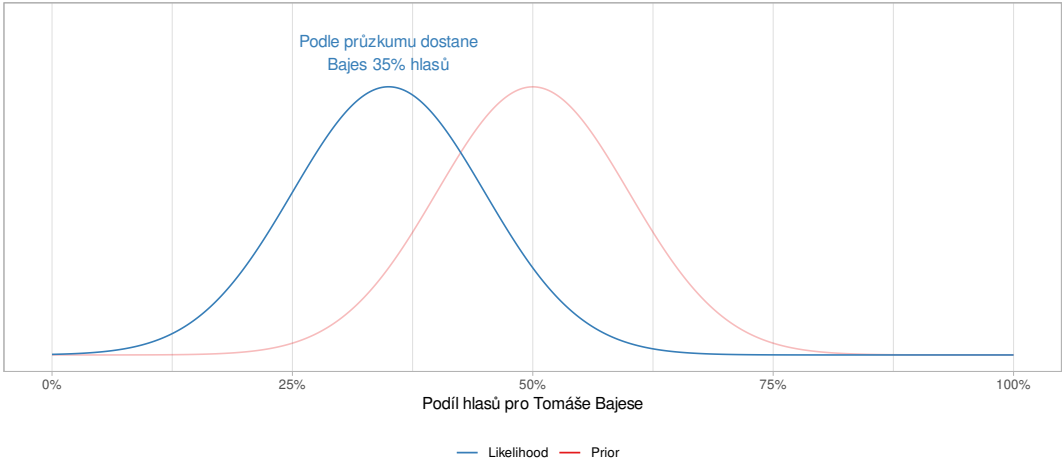
Příklad - Náš prior



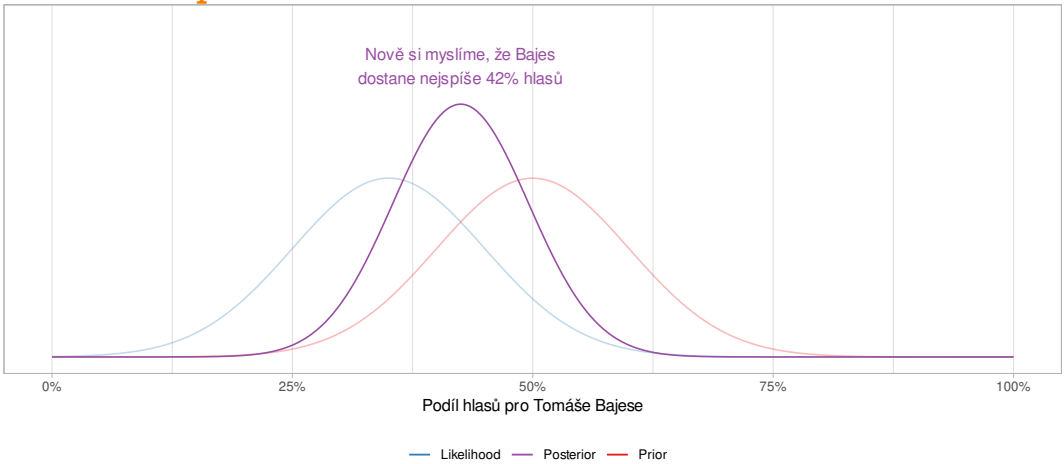
Příklad - Naše likelihood

- Protože toho o novém kandidátovi moc nevíme, tipujeme, že dostane mezi 30% a 70% hlasů (předchozí slide)
- CVVM zveřejnilo nový průzkum, podle kterého má Tomáš Bayes dostat 35 procent hlasů, se standardní chybou 5 procent.
- Tuhle informaci můžeme začlenit do našeho pohledu na svět.

Příklad - Naše likelihood



Příklad - Náš posterior



Priory

Výhody a nevýhody priorů

- Naše priory ovliňují výsledek, musíme je tedy vybírat opatrně
- Na druhou stranu, dobře zvolené priory pomáhají zpřesnit výsledky, snižují nároky na velikost vzorku a zabraňují nesmyslným výsledkům.
- Pro bayesiánskou analýzu musí být nějaký prior zvolen (i kdyby naším priorem bylo, že nic nevíme).

Vliv priorů na výsledek

- Vliv prioru a likelihoodu na posterior nemusí být stejně velký.
- Čím větší vzorek, tím větší vliv likelihood.
- Čím silnější/přesnější prior, tím větší jeho vliv.
- Výběr prioru je tedy důležitý primárně u relativně malých vzorků.

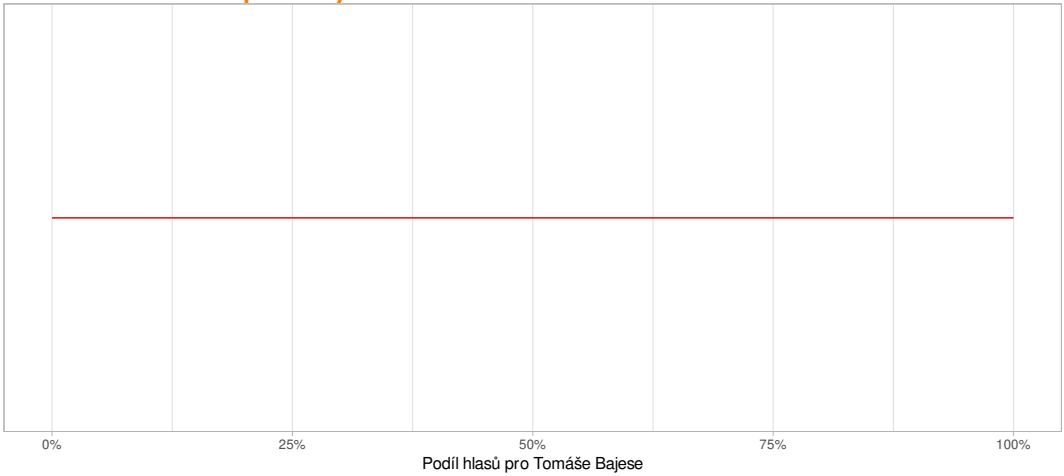
Jak vybírat priory

- Zdrojem prioru může být teoreticky cokoliv (od meta analýzy po expertní úsedek)
- Jediným technickým požadavkem je, že musí jít o pravděpodobnostní rozdělení.
  - "Můj kandidát dostane 50 procent´ hlasů" není prior
  - "Můj kandidát dostane $N(m = 0.5, sd = 0.1)$ procent hlasů" je prior

Typy priorů

- Obecné kategorie priorů
  - Neinformativní priory
  - Slabě informativní priory
  - Silně informativní priory
- Do které kategorie náš prior patří závisí na kontextu.

Neinformativní priory



Neinformativní priory

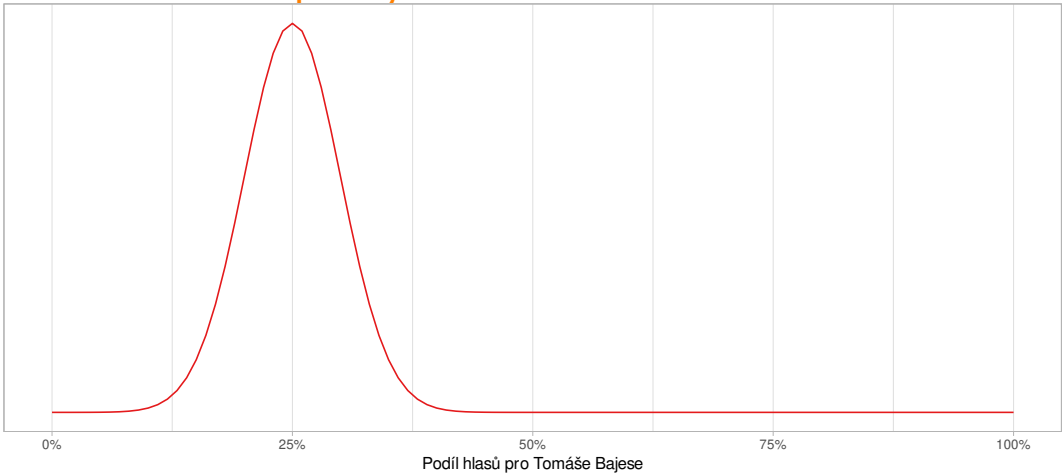
- Vyjadřují, že nic nevíme. Spoléháme pouze na data, která máme k dispozici
  - Tohle je v podstatě frekventistický způsob
- V praxi se nedoporučují
  - Jsou nerealistické
  - Výpočetní problémy

Slabě infomativní priory

Slabě infomativní priory

- Nemají moc velký vliv na výsledek
- Slouží primárně k předcházení výsledků, kterou jsou považovány za extrémně nepravděpodobné
- Většina výzkumníků používá slabě informativní priory

Silně informativní priory



Silně informativní priory

- Mají relativně velký vliv na výsledek.
- Používají se, pokud existuje hodně informací na dané téma nebo pokud je nutné suplovat malý vzorek.
- Nutné velmi dobře teoreticky odůvodnit.

Bayesiánská statistika v R (konečně)

Bayesiánská statistika v R

- Bayesiánské modely zpravidla relativně náročné na výpočet.
- Bayesiánské modely se proto v R nepočítají, protože R je na to moc pomalé.
- Nejčastěji se používá programovací jazyk Stan.
- Naštěstí pro nás, Stan je možné ovládat skrze R

Baličky pro bayesiánskou statistiku

- `rstan` - balíček pro posílání dat mezi Rkem a Stanem
- `rstanarm` - předpřipravené funkce pro urychlení práce, na způsob `lm()` a `glm()`
- `brms` - pokročilejší modely, např. multilevel modely, IRT, apod.

Výpočet bayesiánských modelů

- V praxi 2 způsoby, jak se dopočítat modelu
  - Analyticky (MAP, Grid) - výpočet pomocí vzorce, nuntné počítat hromadu integrálů
  - Pomocí simulae (MCMC) - aproximativní výpočet pomocí tzv. Markov chain Monte Carlo

Markov chain Monte Carlo

- U komplikovaných modelů se nemůžeme dopočítat posteriorního rozdělení přímo, ale můžeme z nich tahat vzorky.
- Výsledkem modelu je vzorek pozorování z posteriou, který můžem sumarizovat.
- Při interpretaci modelu si musíme dát pozor, aby simulace proběhla v pořádku.
- geniálně a graficky vysvětlené zde.

A teď už konečně do Rka...