Past1 – poznámky ze statistiky

25. května 2026

Obsah

Úvod

Budeme se zabývat klasickou statistikou, které se také říká frekvenční statistika. Jiný přístup je bayesovská statistika, která se typicky probírá až později. V těchto poznámkách budeme pracovat hlavně s tímto frekvenčním pohledem: parametr je neznámá, ale pevná hodnota, zatímco náhodná jsou data, která podle daného parametru vznikla.

Statistika je práce s daty. Občas se jí neformálně říká *inverzní pravděpodobnost* – zatímco v klasické teorii pravděpodobnosti obvykle začínáme modelem a ptáme se, jaká data by mohl vygenerovat, ve statistice postupujeme opačně: data už máme a chceme z nich usoudit, jaký model je mohl rozumně vygenerovat.

Budeme se bavit hlavně o třech otázkách:

- **Bodové odhady:** kolik tanků asi vyrobili Němci, když známe jen sériová čísla několika nalezených tanků?
- **Testování hypotéz:** rodí se statisticky prokazatelně víc chlapců než dívek, nebo pozorovaný rozdíl může být náhoda?
- **Intervalové odhady:** co přesně znamenají error bary v grafech?

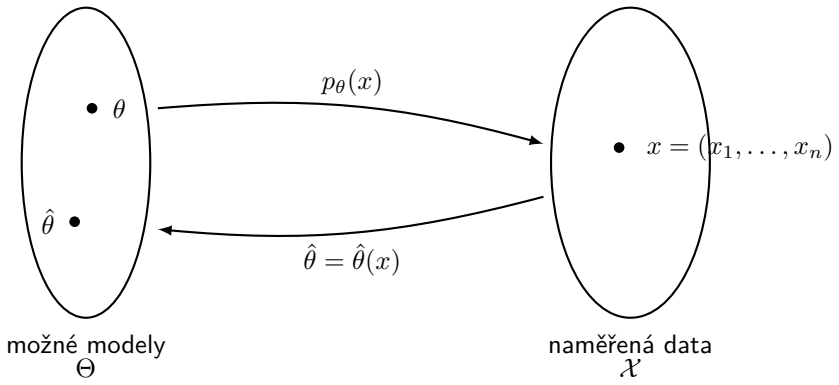
1 Bodové odhady

Teď si vysvětlíme takzvané *bodové odhady*. Bodový odhad je jedno číslo nebo jeden vektor, který z dat vyrobíme jako náš nejlepší odhad neznámého parametru. Například:

- z nalezených sériových čísel tanků odhadujeme celkový počet vyrobených tanků,
- z počtu narozených chlapců a dívek odhadujeme pravděpodobnost narození chlapce,
- z měření výšky a váhy odhadujeme sklon nejlepší přímky, která předpovídá váhu, když známe výšku.

Slovo *bodový* zdůrazňuje, že výsledkem není celý interval možných hodnot, ale jeden konkrétní kandidát. O nejistotě se budeme bavit později.

Obecný rámec pro bodové odhady je následující. Máme prostor možných pozorování $\mathcal{X}$, prostor parametrů Θ a pro každé $\theta \in \Theta$ pravděpodobnostní rozdělení P_θ na $\mathcal{X}$. Parametr θ je neznámá hodnota, kterou se snažíme odhadnout. Každé rozdělení P_θ je pravděpodobnostní model našich dat; naším cílem je najít takový model, tedy takové θ, které na data co nejvíce pasuje.



Obrázek 1: Pravděpodobnostní model vede od parametru θ k rozdělení dat. Statistický odhad jde opačným směrem: z naměřených dat x vyrobí odhad $\hat{\theta}(x)$.

Několik konkrétních příkladů:

- **Německé tanky.** Data jsou nalezená sériová čísla, například $x = \{19, 42, 73\}$. Parametr je celkový počet vyrobených tanků N, takže $\Theta = \{1, 2, 3, \ldots\}$. Pro každé N tedy máme jeden pravděpodobnostní model, který říká, že pozorovaná sériová čísla jsou náhodný výběr z množiny $\{1, \ldots, N\}$.[1]

[1] Notace je ve statistice často složitá, proto ji zde rozepíšeme. O datech přemýšlíme jako o náhodných veličinách $X_1, \ldots, X_n$. Náhodné veličiny mohou nabývat mnoha hodnot a když se bavíme o konkrétní pozorované hodnotě, typicky používáme malá písmena $x_1, \ldots, x_n$. Například pro $N = 100$ je naším modelem pravděpodobnostní rozdělení P_{100}. Pokud pořadí sériových čísel nehraje roli, platí například

$$P_{100}(\{X_1, X_2, X_3\} = \{19, 42, 73\}) = \frac{1}{\binom{100}{3}}.$$

Zápis P_{N_0} znamená, že pravděpodobnost počítáme v modelu s pevně zvolenou hodnotou parametru N_0. Není to podmíněná pravděpodobnost vzhledem k náhodnému jevu $N = N_0$; v klasické statistice zde parametr nebereme jako náhodnou veličinu.

- **Narození chlapce.** Data jsou nuly a jedničky, tedy $x = (x_1, \dots, x_n)$ a $x_i \in \{0, 1\}$. Parametr je pravděpodobnost narození chlapce $p \in [0, 1]$, tedy $\Theta = [0, 1]$. Model říká, že veličiny $X_1, \dots, X_n$ jsou nezávislé a každá má Bernoulliho rozdělení $\mathrm{Bern}(p)$.
- **Normální průměr.** Data jsou reálná čísla $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Předpokládáme, že pochází z normální distribuce $N(\mu, \sigma^2)$, jejíž střed μ neznáme, ale σ ano, tedy $\Theta = \mathbb{R}$.
- **Lineární regrese.** Data jsou dvojice (x_i, y_i). Číslo x_i bereme jako vstup, podle kterého chceme předpovídat y_i. Model říká, že typická hodnota y_i leží přibližně na přímce $a + bx_i$ a rozdíl mezi skutečným pozorováním a touto přímkou je náhodný šum:

  $$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

  kde chyby $\varepsilon_1, \dots, \varepsilon_n$ jsou nezávislé a každá má rozdělení $N(0, 1)$. Odhad potom vrací regresní přímku $y = \hat{a} + \hat{b}x$. V tomto případě je parametrem dvojice čísel, neboli $\Theta = \mathbb{R}^2$.

1.1 Metoda maximální věrohodnosti

1.1.1 Věrohodnost

Základní způsob, jak vyrábět estimátory, je vybrat takovou hodnotu θ, která nejsilněji predikuje data, která jsme opravdu viděli. Tomu se říká metoda maximální věrohodnosti (anglicky *maximum likelihood estimation*, zkratka MLE). Nejprve si pojďme definovat, co je věrohodnost.

Diskrétní případ. Jestliže mají data pravděpodobnostní funkci $p_\theta(x)$[2], věrohodnost (anglicky *likelihood*) parametru θ po pozorování konkrétní hodnoty x definujeme jako

$$L(\theta; x) = p_\theta(x), \qquad \theta \in \Theta.$$

Stejný výraz $p_\theta(x)$ se tedy dá číst dvěma různými způsoby:

- jako funkci x při fixním θ je to model dat, říká, s jakou pravděpodobností uvidíme jaká data,
- jako funkci θ při fixních pozorovaných datech x je to věrohodnost: pro pozorovaná data x říká, jak silně ten který model predikoval data, která jsme nakonec opravdu uviděli.

Spojitý případ. Jestliže data pocházejí ze spojité distribuce, budeme hustotu pořád značit $p_\theta(x)$. Definice likelihoodu je potom stejná a matematika je obdobná diskrétnímu případu.

1.1.2 Definice MLE

Definice. *Odhad metodou maximální věrohodnosti* (anglicky *maximum likelihood estimator*, zkratka MLE) je hodnota parametru, která maximalizuje likelihood pozorovaných dat $x_1, \dots, x_n$:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta \in \Theta} L(\theta; x_1, \dots, x_n).$$

Jinými slovy, MLE filozofie říká, že jako odhad θ máme brát hodnotu, pod kterou jsou pozorovaná data podle zvoleného modelu nejvěrohodnější. V definici tiše předpokládáme, že maximizer existuje a je jednoznačný.

Log-likelihood. Jestliže $X_1, \dots, X_n$ jsou nezávislé náhodné veličiny pocházející ze stejného rozdělení (anglicky *independent and identically distributed*, zkratka iid), pak se likelihood dá napsat takto:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i).$$

Proto se často pracuje s takzvaným log-likelihoodem

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i).$$

[2] V problému německých tanků například $p_{100}(\{19, 42, 73\}) = 1/\binom{100}{3}$.

Logaritmus je rostoucí funkce, takže maximalizace L a maximalizace ℓ jsou tentýž problém. Výhodou log-likelihoodu je, že často vede na jednodušší matematiku, protože součiny mění na součty.

1.1.3 Příklady

Pojďme projít několik statistických problémů a najít v nich MLE. Mnoho běžných úloh ze statistiky a machine learningu jsou speciální případy této logiky.

Praktický recept

U MLE se skoro vždy opakuje stejný postup: zvol model, napiš likelihood, zalogaritmuj, zahoď členy nezávislé na parametru, maximalizuj přes parametr a zkontroluj krajní případy.

Příklad 1: odhad pravděpodobnosti panny

Máme minci, na které padne panna s pravděpodobností p, a tuto pravděpodobnost chceme odhadnout. Hodíme proto mincí n-krát. Výsledky modelujeme jako nezávislé náhodné veličiny

$$X_1, \dots, X_n,$$

kde každá má rozdělení $\mathrm{Bern}(p)$. Hodnota $X_i = 1$ znamená, že v i-tém hodu padla panna, a $X_i = 0$ znamená opak. Pozorujeme konkrétní hodnoty $x_1, \dots, x_n \in \{0, 1\}$ a označíme

$$k = \sum_{i=1}^{n} x_i,$$

tedy počet hodů, ve kterých padla panna.

Prostor parametrů je $\Theta = [0, 1]$ a neznámým parametrem je přímo p. Chceme najít estimátor $\hat{p}_{\mathrm{MLE}}$, který maximalizuje pravděpodobnost pozorovaných dat $x_1, \dots, x_n$. Likelihood pro dané p je

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = p^k(1-p)^{n-k}$$

Log-likelihood je tedy

$$\ell(p) = \log L(p) = k \log p + (n-k)\log(1-p).$$

Pro $0 < k < n$ leží maximum uvnitř intervalu $(0, 1)$, takže ho najdeme derivací. Spočteme

$$\ell'(p) = \frac{k}{p} - \frac{n-k}{1-p}.$$

Z rovnice $\ell'(p) = 0$ dostaneme

$$\frac{k}{p} = \frac{n-k}{1-p},$$

a tedy

$$k(1-p) = (n-k)p.$$

Po úpravě vyjde

$$\hat{p}_{\mathrm{MLE}} = \frac{k}{n}.$$

Stejný vzorec dává správný výsledek i v krajních případech: pokud $k = 0$, maximum je v $p = 0$, a pokud $k = n$, maximum je v $p = 1$.

Interpretace

MLE zde odvodí přirozený estimátor pravděpodobnosti panny: pozorovaný podíl hodů, ve kterých padla panna.

Příklad 2: normální model se známým rozptylem

Naměřili jsme IQ n matfyzáků. Jednoduchý model je, že pozorování pocházejí z normálního rozdělení:

$$X_1, \dots, X_n,$$

kde pozorování jsou nezávislá a každé má rozdělení $N(\mu, \sigma^2)$. Parametr $\mu \in \mathbb{R}$ je neznámý a chceme ho odhadnout. Směrodatnou odchylku σ, a tedy i rozptyl σ^2, budeme pro jednoduchost považovat za známé; například $\sigma = 15$. Pozorovaná data označíme $x_1, \dots, x_n$.

Pro dané μ si nejprve napíšeme hustotu jednoho pozorování:

$$p_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Likelihood celého vzorku je tedy

$$L(\mu) = \prod_{i=1}^{n} p_\mu(x_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Po zalogaritmování dostaneme

$$\ell(\mu) = \log L(\mu) = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

První člen nezávisí na μ a kladný faktor $1/(2\sigma^2)$ také nemění polohu maxima. Maximalizace $\ell(\mu)$ je proto totéž jako minimalizace součtu čtvercových odchylek

$$Q(\mu) = \sum_{i=1}^{n}(x_i-\mu)^2.$$

Derivace je

$$Q'(\mu) = -2\sum_{i=1}^{n}(x_i-\mu).$$

Rovnice $Q'(\mu) = 0$ dává

$$\sum_{i=1}^{n}(x_i-\mu) = 0,$$

a tedy

$$\sum_{i=1}^{n} x_i - n\mu = 0.$$

Pro pozorovaná data tedy vychází

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Jako estimator, tedy jako funkci náhodných dat, píšeme

$$\hat{\mu}_{\mathrm{MLE}} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Druhá derivace $Q''(\mu) = 2n > 0$ říká, že Q má minimum, takže log-likelihood má maximum.

Interpretace

MLE zde opět vede na přirozený vzorec: střed normálního rozdělení odhadneme výběrovým průměrem. Důležité je, že průměr jsme do zadání nevložili ručně; vyšel z normálního modelu a z maximalizace likelihoodu. Stejná logika stojí za metodou nejmenších čtverců.

Příklad 3: problém německých tanků

Během druhé světové války se Spojenci snažili odhadnout, kolik kusů techniky Německo vyrábí. Jedna z užitečných stop byla překvapivě obyčejná: sériová čísla na ukořistěných nebo zničených kusech. Z několika pozorovaných čísel se dá statisticky odhadovat celkový počet vyrobených kusů. Tento příklad je známý jako German tank problem.

V modelu předpokládáme, že existuje neznámé celé číslo N a vyrobené kusy mají sériová čísla

$$\{1, \dots, N\}.$$

Pozorujeme k různých sériových čísel vybraných bez vracení. Parametrický prostor je

$$\Theta = \{k, k+1, k+2, \dots\},$$

protože celkový počet kusů musí být alespoň počet pozorovaných kusů. Parametr je konkrétní, ale neznámá hodnota $N \in \Theta$. Estimator, například $\hat{N}$, je naopak funkce dat.

Označme

$$m = \max(x_1, \dots, x_k)$$

největší pozorované sériové číslo. Při fixním N má každá k-prvková podmnožina stejnou pravděpodobnost

$$\frac{1}{\binom{N}{k}}.$$

Likelihood je proto

$$L(N) = \begin{cases} \binom{N}{k}^{-1}, & N \geq m, \\ 0, & N < m. \end{cases}$$

Pro $N \geq m$ roste $\binom{N}{k}$ s N, takže $\binom{N}{k}^{-1}$ klesá. Maximum je tedy v nejmenší hodnotě kompatibilní s daty:

$$\hat{N}_{\text{MLE}} = m.$$

Jde o maximum na hranici prostoru parametrů: data pouze říkají, že musí platit $N \geq m$, a likelihood je největší v nejmenší takové hodnotě.

Pro pozorování

$$19,\ 40,\ 42,\ 60$$

tedy MLE vychází

$$\hat{N}_{\text{MLE}} = 60.$$

Příklad 4: lineární regrese jako MLE

Lineární regrese je o tom, že jednou nebo více vstupními proměnnými predikujeme číselnou veličinu. Například můžeme chtít odhadovat hmotnost člověka z obvodu pasu. Na obrázku je každý bod jeden člověk a přímka je lineární model, který zachycuje hlavní trend v datech.

Začneme nejjednodušší verzí s jednou vstupní proměnnou. Pro každé pozorování máme vstup x_i a výstup y_i. Hledáme přímku

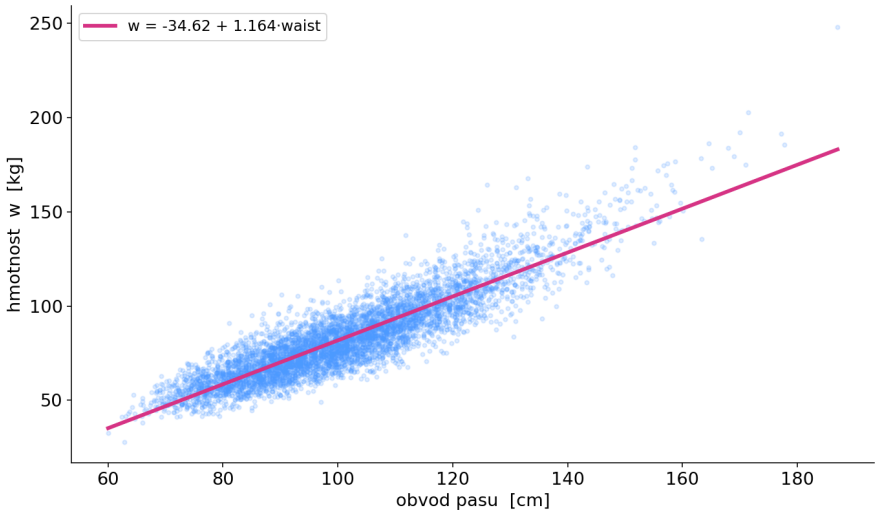
$$y = a + bx,$$

která data dobře vystihuje. Metoda nejmenších čtverců vybírá takové a a b, které minimalizují

$$\text{RSS}(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Teď ukážeme, že tento vzorec není náhodný trik, ale MLE v normálním modelu. Pro jednoduchost položme $\sigma = 1$ a předpokládejme

$$Y_i = a + bx_i + \varepsilon_i, \qquad i = 1, \dots, n.$$



Obrázek 2: Jednoduchá lineární regrese: hmotnost jako funkce obvodu pasu.

Člen ε_i je chyba modelu: rozdíl mezi skutečnou hodnotou Y_i a hodnotou $a+bx_i$, kterou by předpověděla přesná přímka. Předpokládáme, že chyby $\varepsilon_1, \ldots, \varepsilon_n$ jsou nezávislé, mají nulovou střední hodnotu a každá má normální rozdělení $N(0,1)$. Vstupy x_i zde bereme jako známé naměřené hodnoty, podle kterých predikujeme Y_i.

Likelihood pro parametry (a, b) je

$$L(a,b) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - a - bx_i)^2}{2}\right).$$

Log-likelihood je

$$\ell(a,b) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

První člen je konstanta a faktor $1/2$ nemění polohu maxima. Maximalizace log-likelihoodu podle (a, b) je proto totéž jako minimalizace reziduálního součtu čtverců

$$\mathrm{RSS}(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

Z derivace lze dostat explicitní vzorce. Pokud nejsou všechna x_i stejná, pak

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Tyto vzorce teď nemusíme dále rozebírat; podstatné je, že existují a vznikají přesně z minimalizace $\mathrm{RSS}(a,b)$.

Pointa lineární regrese

Metoda nejmenších čtverců je MLE estimátor, který předpokládá, že šum v datech má normální rozdělení.

Obecná lineární regrese může mít mnoho vstupních proměnných, nejen jednu; díky tomu je velmi univerzální a používá se jako základní model i jako stavební blok složitějších metod.

Příklad 5: klasifikace a cross-entropy

Představme si, že trénujeme neuronovou síť (třeba AlexNetem) na rozpoznávání obrázků z nějakého datasetu (třeba CIFAR-10). Každý obrázek patří do jedné z deseti tříd (letadlo, auto, pták, ...). Model, třeba neuronová síť, vezme obrázek x a vrátí pravděpodobnosti jednotlivých tříd. Píšeme

$$p_\theta(y \mid x),$$

kde θ jsou váhy modelu a y je třída. Optimalizaci neuronové síťe teď můžeme chápat jako speciální případ bodového odhadu. Pro trénovací data (x_i, y_i) je likelihood sítě θ následující:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(y_i \mid x_i).$$

Log-likelihood je tedy

$$\ell(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i \mid x_i).$$

Maximalizovat log-likelihood je totéž jako minimalizovat negativní log-likelihood

$$-\ell(\theta) = -\sum_{i=1}^{n} \log p_\theta(y_i \mid x_i).$$

V machine learningu se tomuto cíli často říká *cross-entropy loss*. Z hlediska statistiky optimalizace crossentropie odpovídá MLE, neboli hledání modelu, který nejsilněji predikuje pozorovaná data.

Další příklady

Model	MLE výsledek a poznámka
Normální model s neznámým μ i σ^2	$\hat{\mu}_{\text{MLE}} = \bar{X}, \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_i (X_i - \bar{X})^2$.
Poissonovo rozdělení Pois(λ)	$\hat{\lambda}_{\text{MLE}} = \bar{X}$. Stejně jako u normálního průměru vyjde pozorovaný průměr.
Exponenciální rozdělení Exp(λ)	$\hat{\lambda}_{\text{MLE}} = 1/\bar{X}$. Parametr λ je intenzita; větší průměrná doba znamená menší intenzitu.

Na přednáškách z machine-learningu se můžete setkat s následujícími technikami: logistická regrese, k-means, PCA, nebo matrix factorization. Všechny tyto techniky se víceméně dají interpretovat jako MLE estimátory.

1.2 Kvalita odhadů: vychýlení, variance a MSE

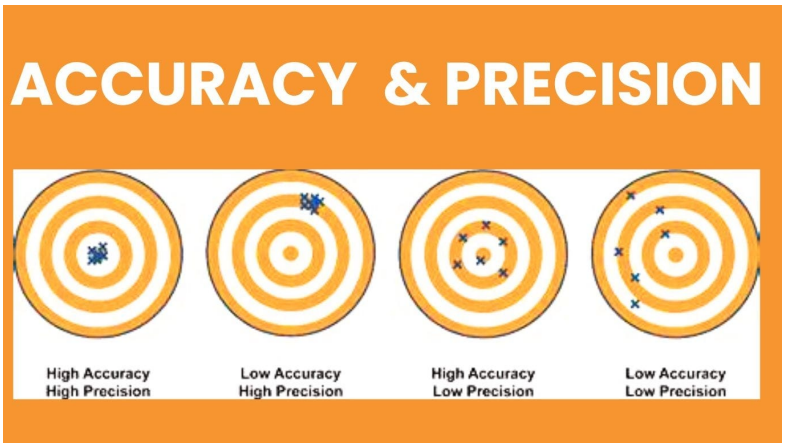
Po odvození MLE ještě nevíme, zda je odhad dobrý. Odhad může být systematicky posunutý, může hodně kolísat mezi opakovanými vzorky, nebo obojí. V angličtině se tyto dvě vlastnosti často popisují slovy *accuracy* a *precision* a kreslí se pomocí terče: accuracy říká, zda míříme na správné místo, precision říká, jak těsně jsou zásahy u sebe.

Ve statistickém jazyce budeme systematický posun měřit pomocí *vychýlení* (anglicky *bias*) a kolísání mezi vzorky pomocí *variance estimatoru* (anglicky *variance of an estimator*).

Rozlišovat bias a varianci se vyplatí i v běžném životě. Pokud nás zajímá přesnost volebního průzkumu na obrázku níže, ptáme se na dvě oddělené otázky. Jde o voličský potenciál, takže uvedená čísla se nemusí sčítat na 100%.

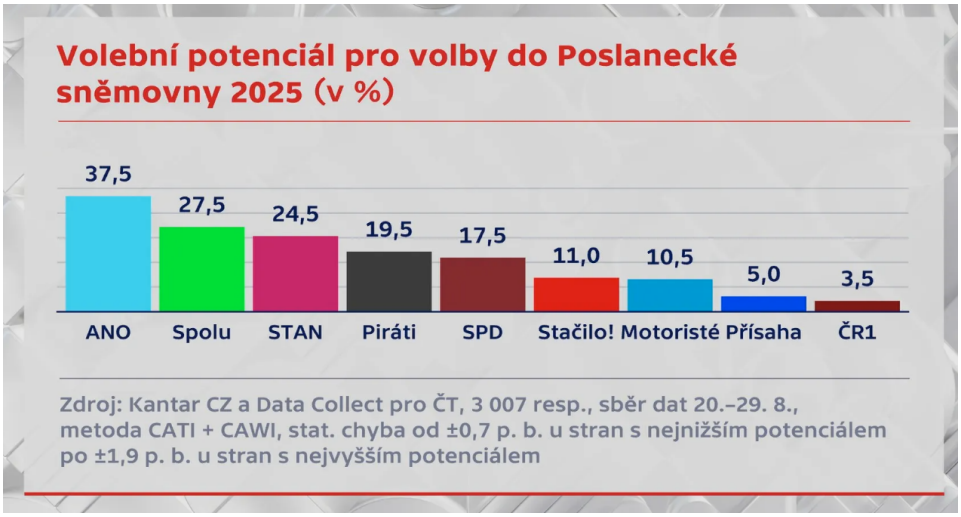
Jaký je bias průzkumu? Zkratka CATI znamená telefonické dotazování, CAWI dotazování přes internet. Takové metody mohou systematicky podhodnocovat skupiny, které telefon nebo internet používají jinak než zbytek populace. Vychýlení se proto v praxi řeší následným zpracováním dat, například převážením odpovědí podle věku, vzdělání nebo regionu.

Jaká je variance průzkumu? Pokud se zeptáme několika tisíc respondentů, směrodatná chyba odhadu podílu bývá řádově jednotky procentních bodů. Přesněji, pro jednoduchý odhad podílu $\hat{p}$ níže



Obrázek 3: Systematický posun a rozptyl mezi opakovanými vzorky jsou dvě různé vlastnosti estimatoru.

dostaneme $\mathrm{Var}(\hat{p}) = p(1-p)/n$, tedy směrodatnou chybu řádu $1/\sqrt{n}$. Pokud chceme směrodatnou chybu zmenšit desetkrát, musíme se zeptat zhruba stokrát více respondentů. Proto se v praxi často nevyplatí zvyšovat velikost průzkumu daleko za několik tisíc lidí; přesnost se pak zlepšuje už jen pomalu.



Obrázek 4: Volební průzkum: vychýlení souvisí s výběrem respondentů, variance s počtem odpovědí.

1.2.1 Vychýlení

Estimator je funkce dat

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n),$$

a proto je sám náhodnou veličinou. Kdybychom experiment opakovali, dostávali bychom různé hodnoty $\hat{\theta}$. Kvalita estimatoru popisuje právě toto rozdělení přes opakované vzorky.

Definice. *Vychýlení* estimatoru $\hat{\theta}$ parametru θ (anglicky *bias*) je

$$\mathrm{bias}_\theta(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta.$$

Estimator je *nestranný* (anglicky *unbiased*), jestliže

$$\mathbb{E}_\theta[\hat{\theta}] = \theta$$

pro každou hodnotu θ v prostoru parametrů.

Vychýlení je systematický posun. Neříká, jak moc estimator kolísá mezi vzorky; říká, kam míří v průměru. Pokud je estimator vychýlený, někdy z něj umíme vyrobit nestranný estimator jednoduchou korekcí, například vynásobením vhodnou konstantou nebo odečtením známého posunu. Nestrannost ale není unikátní vlastnost, která by určovala jediný správný odhad. Pokud máme nestranný estimator a začneme ho počítat jen z první poloviny dat, zatímco druhou polovinu zahodíme, pořád může zůstat nestranný. Bude však typicky méně přesný, protože používá méně informací.

1.2.2 Příklady

Testování mince.

V příkladu s mincí jsme dostali MLE

$$\hat{p} = \frac{k}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Jestliže jsou $X_1, \ldots, X_n$ nezávislé a každá veličina má rozdělení $\mathrm{Bern}(p)$, pak $\mathbb{E}[X_i] = p$. Linearita střední hodnoty proto dává

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p.$$

Odhad $\hat{p} = k/n$ je tedy nestranný estimator pravděpodobnosti panny. Navíc

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{p(1-p)}{n},$$

protože $\mathrm{Var}(X_i) = p(1-p)$. Směrodatná chyba odhadu podílu tedy klesá jako $1/\sqrt{n}$.

Vychýlení u německých tanků.

U problému německých tanků jsme jako MLE dostali maximum pozorovaných sériových čísel. Intuitivně čekáme, že takový odhad bude podstřelovat: pokud nevidíme všechny vyrobené kusy, největší nalezené číslo typicky nebude úplně poslední vyrobené číslo. Spočítáme to přesně.

Jako náhodnou veličinu označme maximum

$$M = \max(X_1, \ldots, X_k).$$

Pokud je skutečný počet kusů N, pak

$$\mathbb{P}_N(M = r) = \frac{\binom{r-1}{k-1}}{\binom{N}{k}}, \qquad r = k, \ldots, N.$$

Aby totiž bylo maximum rovno r, musí být číslo r ve vzorku a z předchozích $r-1$ čísel vybereme zbývajících $k-1$.

Z identity

$$r\binom{r-1}{k-1} = k\binom{r}{k}$$

a hockey-stick identity

$$\sum_{r=k}^{N}\binom{r}{k} = \binom{N+1}{k+1}$$

dostaneme

$$\mathbb{E}_N[M] = \sum_{r=k}^{N} r\frac{\binom{r-1}{k-1}}{\binom{N}{k}} = \frac{k}{\binom{N}{k}}\sum_{r=k}^{N}\binom{r}{k} = \frac{k\binom{N+1}{k+1}}{\binom{N}{k}} = \frac{k(N+1)}{k+1}.$$

Vychýlení estimatoru M jako odhadu N je tedy

$$\mathrm{bias}_N(M) = \mathbb{E}_N[M] - N = \frac{k-N}{k+1}.$$

Pro $N > k$ je záporné: maximum pozorovaných sériových čísel typicky podstřelí skutečný počet kusů.

Z tohoto výpočtu lze vyrobit nestrannou korekci

$$\hat{N}_{\mathrm{unb}} = \frac{k+1}{k} M - 1,$$

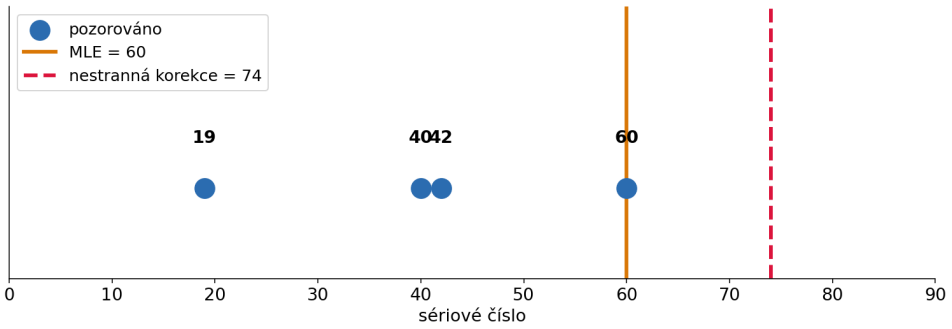
protože

$$\mathbb{E}_N[\hat{N}_{\mathrm{unb}}] = \frac{k+1}{k}\mathbb{E}_N[M] - 1 = N.$$

Pro pozorování $19, 40, 42, 60$ máme $k = 4$ a $M = 60$, takže

$$\hat{N}_{\mathrm{unb}} = \frac{5}{4} \cdot 60 - 1 = 74.$$

MLE tedy vychází 60, zatímco nestranně korigovaný odhad vychází 74. To není spor: maximální věrohodnost a nestrannost jsou dvě různá kritéria.



Obrázek 5: Německé tanky: MLE bere největší pozorované sériové číslo, nestranná korekce ho posune výš.

1.2.3 Variance a směrodatná chyba

Definice. *Variance estimatoru* (anglicky *variance of an estimator*) je

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2].$$

Směrodatná chyba estimatoru (anglicky *standard error*) je směrodatná odchylka estimatoru:

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{Var}_\theta(\hat{\theta})}.$$

Příklad: standard error podílu a průměru.
U falešné mince máme $X_i \sim \mathrm{Bernoulli}(p)$ a estimator

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}.$$

Protože $\mathrm{Var}(X_i) = p(1-p)$, dostáváme

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}, \qquad \mathrm{se}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

MLE $\hat{p} = k/n$ je tedy sice jedna konkrétní hodnota spočtená z dat, ale kdybychom celý experiment opakovali, kolísal by kolem skutečného p typicky v měřítku $\sqrt{p(1-p)/n}$.

Podobně pro výběrový průměr nezávislých stejně rozdělených veličin s variancí σ^2 platí

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\sigma^2}{n}.$$

Tedy

$$\mathrm{se}(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

V normálním modelu, kde jsou $X_1, \ldots, X_n$ nezávislé a každá veličina má rozdělení $N(\mu, \sigma^2)$ se známým σ, je tedy standard error estimatoru $\hat{\mu} = \bar{X}$ přesně $\sigma/\sqrt{n}$.

Odmocninové pravidlo

Přidávání dat pomáhá, ale ne lineárně. Chceme-li zhruba dvakrát menší typickou chybu průměru nebo podílu, potřebujeme asi čtyřikrát větší vzorek; pro desetkrát menší chybu potřebujeme asi stokrát větší vzorek.

1.2.4 Střední kvadratická chyba

Definice. *Střední kvadratická chyba* estimatoru $\hat{\theta}$ (anglicky *mean squared error*, zkratka MSE) je

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \theta)^2].$$

MSE kombinuje systematickou chybu i náhodné kolísání. Přesný vztah je

Věta: rozklad MSE

Pro každý estimator $\hat{\theta}$ a každé θ, pro které existují příslušné momenty, platí

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{bias}_\theta(\hat{\theta})^2 + \mathrm{Var}_\theta(\hat{\theta}).$$

Důkaz rozkladu MSE. Přičteme a odečteme $\mathbb{E}_\theta[\hat{\theta}]$:

$$\hat{\theta} - \theta = (\hat{\theta} - \mathbb{E}_\theta\hat{\theta}) + (\mathbb{E}_\theta\hat{\theta} - \theta).$$

Po umocnění dostaneme

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2 + 2(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})(\mathbb{E}_\theta\hat{\theta} - \theta) + (\mathbb{E}_\theta\hat{\theta} - \theta)^2.$$

Vezmeme střední hodnotu. Prostřední člen zmizí, protože

$$\mathbb{E}_\theta[\hat{\theta} - \mathbb{E}_\theta\hat{\theta}] = 0.$$

Zbude

$$\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta\hat{\theta})^2] + (\mathbb{E}_\theta\hat{\theta} - \theta)^2,$$

tedy

$$\mathrm{MSE}_\theta(\hat{\theta}) = \mathrm{bias}_\theta(\hat{\theta})^2 + \mathrm{Var}_\theta(\hat{\theta}).$$

□

Když chceme porovnávat dva estimátory stejné veličiny, MSE je často přirozená metrika: měří průměrnou kvadratickou vzdálenost od pravé hodnoty. Díky rozkladu výše přesně vidíme, jestli estimator prohrává proto, že míří systematicky vedle, nebo proto, že příliš kolísá.

Ale pozor, dva estimátory jsou obecně neporovnatelné. Např. pokud v problému německých tanků uvažujeme MLE estimátor, a estimátor $\hat{N} = 42$, tak druhý estimátor má nulový MSE pokud náhodou platí $N = 42$. Nedá se tedy říct, že by druhý estimátor byl horší. Z hlediska MSE je nicméně druhý estimátor horší skoro pro všechny možné hodnoty N.

2 Testování hypotéz

V testování hypotéz se ptáme, zda jsou data ještě slučitelná s nějakým výchozím vysvětlením, nebo zda už vypadají jako důkaz pro něco zajímavějšího. Výchozí vysvětlení nazýváme *nulová hypotéza* a značíme ho H_0. Typicky říká "nic zajímavého se neděje": mince je férová, nový lék nemá efekt, dvě veličiny spolu nesouvisí. Proti němu stojí *alternativní hypotéza* H_1, která stojí pro všechny ostatní vysvětlení.

Statistický test není nástroj, který přímo rozhodne, co je pravda. Je to pravidlo, které má kontrolovat falešné poplachy: když H_0 ve skutečnosti platí, test ji smí zamítat jen zřídka. To je podstatně jemnější garance, než jak si ji lidé často představujou. Test může zamítnout pravdivou nulovou hypotézu; hladina testu jen říká, jak často se to smí stávat při opakovaném používání stejného postupu.

při $\alpha = 0.05$
asi 1 falešné zamítnutí
z 20 opakování

Obrázek 6: Hladina $\alpha = 0.05$ kontroluje dlouhodobý podíl falešných pozitivních závěrů, ne pravděpodobnost pravdivosti konkrétní hypotézy po pozorování dat.

Často se bavíme o takzvaných parametrických testech (uvidíme z-test, t-test), kde si omezíme množinu všech hypotéz tak, že jsou kontrolovány jen jedním nebo několika čísly. Například předpokládáme, že data pocházejí z normálního rozdělení a my jen nevíme hodnotu μ, σ. Opakem jsou neparametrické testy (uvidíme permutační test).

Definice. *Statistický test* (anglicky *statistical test*) je formálně pravidlo, které na základě dat rozhoduje, zda zamítnout nulovou hypotézu H_0 ve prospěch alternativy H_1.

Formálně lze test zapsat pomocí testovací funkce

$$\varphi : \mathcal{X} \to \{\text{zamítáme}, \text{nezamítáme}\}.$$

Hodnota $\varphi(x) = \text{zamítáme}$ znamená, že při datech x zamítáme H_0. Hodnota $\varphi(x) = \text{nezamítáme}$ znamená, že H_0 nezamítáme. Množina

$$R = \{x \in \mathcal{X} : \varphi(x) = \text{zamítáme}\}$$

se nazývá *kritická oblast* (anglicky *critical region* nebo *rejection region*). Je to množina dat, pro která test zamítá.

2.1 Testová statistika a extrémní hodnoty

Definice. *Testová statistika* (anglicky *test statistic*) je funkce dat, která shrnuje informaci relevantní pro rozhodnutí mezi H_0 a H_1:

$$T = T(X_1, \ldots, X_n).$$

Test typicky zamítá pro extrémní hodnoty T. Co znamená "extrémní" se liší. U jednostranného testu nás zajímá jeden směr, například velké kladné hodnoty T. U oboustranného testu nás zajímá odchylka oběma směry, například velké hodnoty

$$|T|.$$

Když později budeme mluvit o datech "takto extrémních nebo extrémnějších", vždy tím myslíme extrémnost vzhledem ke zvolené alternativě a zvolené testové statistice.

2.2 False positive, false negative a hladina

Test může udělat dva typy chyb:

- **false positive**: zamítneme H_0, i když H_0 platí,
- **false negative**: nezamítneme H_0, i když platí alternativa.

V klasické statistické terminologii se false positive říká také chyba prvního druhu a false negative chyba druhého druhu.

Hladina významnosti (anglicky *significance level*) α je horní mez na pravděpodobnost false positive.

Formálně test φ má hladinu nejvýše α, jestliže

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\text{zamítneme } H_0) = \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\varphi(X) = \text{zamítáme}) \leq \alpha.$$

U jednoduché nulové hypotézy $H_0 : \theta = \theta_0$ je to jen

$$\mathbb{P}_{\theta_0}(X \in R) \leq \alpha.$$

Skutečné maximum

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(\varphi(X) = \text{zamítáme})$$

se někdy nazývá skutečná velikost testu (anglicky *size*). V praxi se slovo hladina často používá i pro cílovou hodnotu, například $\alpha = 0.05$.

Definice. *P-hodnota* (anglicky *p-value*) je nejmenší hladina významnosti, na které by daný test ještě zamítl nulovou hypotézu pro pozorovaná data. Jinými slovy, hypotézu na hladině 0.05 zamítáme tehdy, když nám vyjde p-hodnota menší než 0.05. Místo "zamítáme na hladině 0.05" se tak často v článcích píše p-hodnota, která je informativnější.[3]

Jak p-hodnotu číst správně

P-hodnota není pravděpodobnost, že H_0 je pravdivá. Je to pravděpodobnost takto extrémních nebo extrémnějších dat za předpokladu, že H_0 platí.

2.3 Příklady

Podíl chlapců mezi novorozenci

Představme si, že máme data o narozeních za rok 2022 a chceme vědět, zda jsou konzistentní s jednoduchým modelem 50/50, nebo zda se rodí více chlapců. Pokud přicházíme k datům s předem danou hypotézou "slyšel jsem, že chlapců se rodí trochu víc", dává smysl jednostranný test

$$H_0 : p = 0.5, \qquad H_1 : p > 0.5.$$

Jednostranný test soustředí celou hladinu do jednoho směru, a proto má větší šanci zamítnout, když je skutečná odchylka právě tímto směrem.

Pokud jsme ale žádnou směrovou hypotézu předem neměli a až po pohledu na data vidíme, že chlapců je víc než dívek, korektní je oboustranný test

$$H_0 : p = 0.5, \qquad H_1 : p \neq 0.5.$$

Obecně má být test navržen před pohledem na data. Rozhodnout se pro jednostranný test až poté, co jsme uviděli směr odchylky, je forma p-hackingu.

Jako statistiku vezmeme počet chlapců

$$S = \#\text{chlapců}$$

[3] K p-hodnotám, které těsně nevyšly jako signifikantní, viz například blogový text Still Not Significant 2.

nebo ekvivalentně podíl

$$\hat{p} = \frac{S}{n}.$$

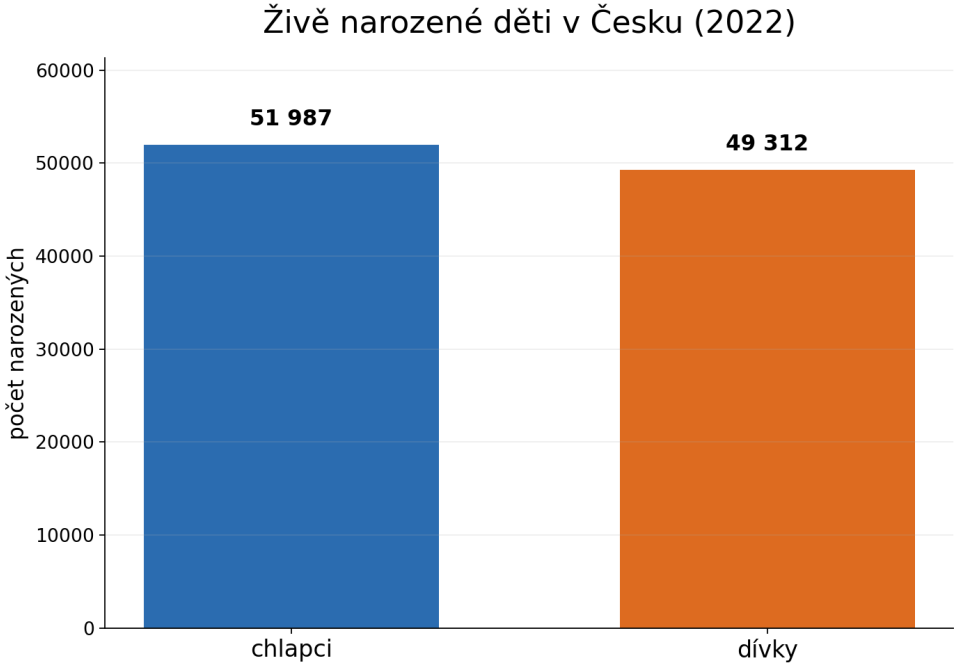
V našich datech vycházelo

$$n = 101299, \qquad \hat{p}_{\text{obs}} = 0.5132.$$

Pod nulovou hypotézou má počet chlapců rozdělení

$$S \sim \text{Bin}(101299, 0.5).$$

Nulové rozdělení můžeme získat dvěma způsoby. Buď ho nasimulujeme: opakovaně vygenerujeme 101299 porodů s pravděpodobností chlapce 0.5 a díváme se, jak často vyjde stejně velká odchylka. Nebo použijeme vzorec pro binomické rozdělení; u tak velkého n ho navíc velmi dobře aproximuje normální rozdělení.



Obrázek 7: Pozorovaná data sama o sobě nestačí. Potřebujeme je porovnat s tím, co bychom čekali pod H_0.

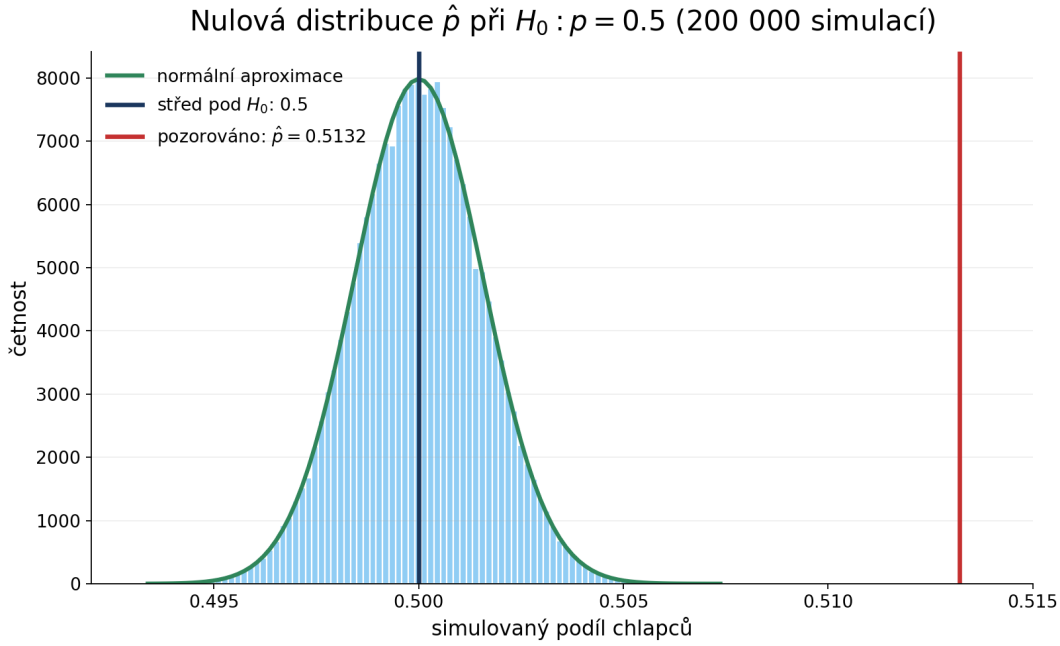
Zamítací oblast

Zamítací oblast je část prostoru možných dat, ve které test zamítá H_0. V oboustranném testu ji při hladině $\alpha = 0.05$ obvykle rozdělíme do dvou krajů nulového rozdělení, po 2.5% na každé straně. Důležité je pořadí: zamítací oblast konstruujeme před pohledem na konkrétní data. Po pozorování už jen zkontrolujeme, zda pozorovaná hodnota do této oblasti spadla.

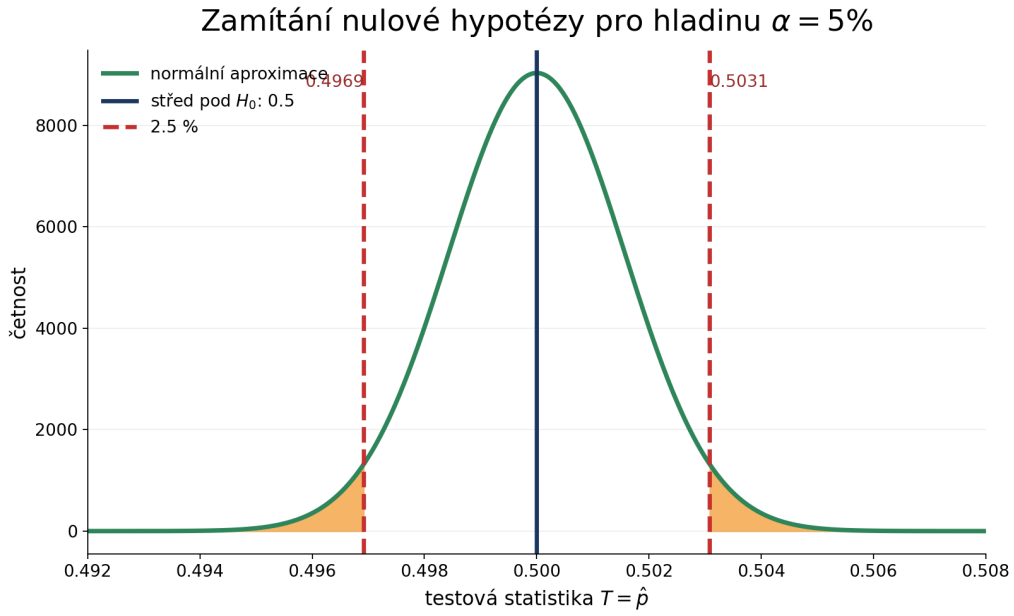
Pro oboustranný test počítáme pravděpodobnost stejně nebo více extrémního výsledku v obou směrech od 0.5. Pro naše data vyšla p-hodnota přibližně

$$p \approx 4.39 \times 10^{-17}.$$

U předem zvoleného jednostranného testu $H_1 : p > 0.5$ by p-hodnota byla přibližně poloviční. Tak malá p-hodnota znamená, že kdyby byl skutečný podíl chlapců přesně 0.5, takto odchýlený výsledek bychom viděli extrémně zřídka. Na běžných hladinách, například 0.05 nebo 0.01, proto H_0 s přehledem zamítáme.



Obrázek 8: Rozdělení testové statistiky pod nulovou hypotézou.



Obrázek 9: Pro hladinu $\alpha = 5\%$ zamítáme H_0 v krajních oblastech nulového rozdělení.

Příklad: z-test průměru při známém rozptylu

Nechť

$$X_1, \dots, X_n$$

jsou nezávislé a každá veličina má rozdělení $N(\mu, \sigma^2)$, kde σ je známé. Testujeme

$$H_0 : \mu = \mu_0 \qquad \text{proti} \qquad H_1 : \mu \neq \mu_0.$$

Za platnosti H_0 víme, že

$$\bar{X} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Standardizovaná testová statistika z-testu je

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Za H_0 má přesně rozdělení

$$Z \sim N(0, 1).$$

Oboustranný test na hladině α zamítá, když

$$|Z| > z_{1-\alpha/2},$$

kde $z_{1-\alpha/2}$ je příslušný kvantil standardního normálního rozdělení. Pro $\alpha = 0.05$ je

$$z_{0.975} \approx 1.96.$$

P-hodnota je

$$p = 2(1 - \Phi(|z_{\text{obs}}|)).$$

2.4 Příklad: t-test při neznámém rozptylu

Když σ neznáme, nahradíme ho výběrovou směrodatnou odchylkou

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Testová statistika je

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

Za předpokladu normálního modelu a platnosti H_0 má statistika t-testu Studentovo rozdělení

$$T \sim t_{n-1}.$$

Tento výsledek není z definice z-testu zřejmý. V čitateli stále stojí odchylka průměru od μ_0, ale jmenovatel už není pevné číslo $\sigma/\sqrt{n}$: odhadujeme ho z dat pomocí $s/\sqrt{n}$. Proto se do statistiky přidává další náhodnost a výsledné rozdělení má těžší ocasy.

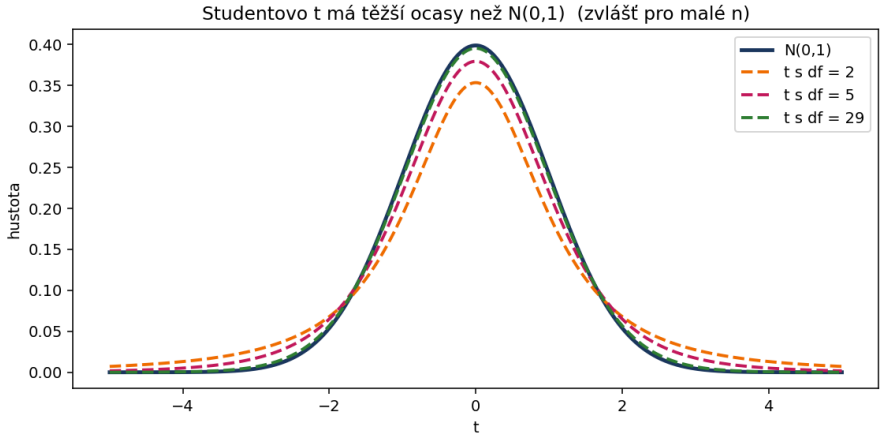
Oboustranný test na hladině α zamítá, když

$$|T| > t_{1-\alpha/2,n-1}.$$

Pro malé vzorky jsou t-kvantily větší než normální kvantily, protože navíc odhadujeme rozptyl. Pro velké n se t-rozdělení blíží standardnímu normálnímu rozdělení.

2.5 Permutační test jako model nulové hypotézy

Ne všechny testy musí stát na normálním modelu. U permutačního testu je nulová hypotéza často tvrzení, že přiřazení štítků je vůči pozorovaným hodnotám náhodné.



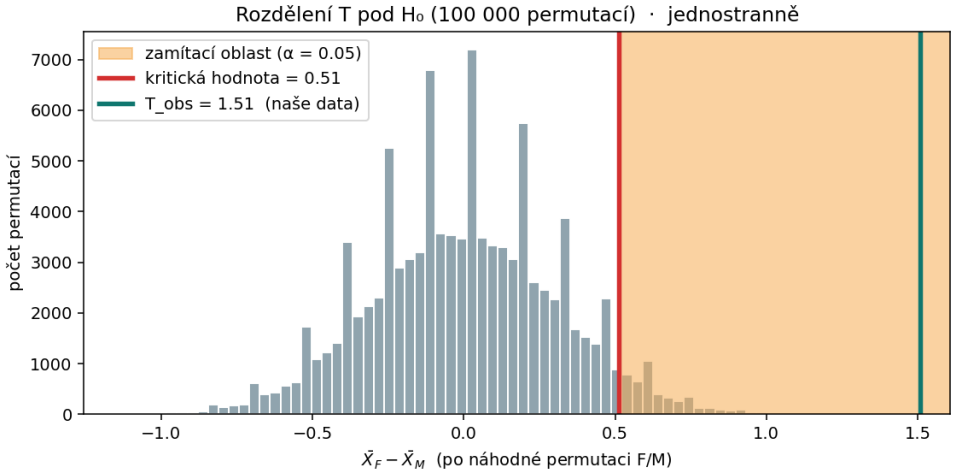
Obrázek 10: Studentovo t-rozdělení má při malém počtu stupňů volnosti těžší ocasy než $N(0,1)$. S rostoucím n se k normálnímu rozdělení přibližuje.

Příklad: délka příjmení podle pohlaví.

Ve třídních datech se můžeme ptát, zda mají ženy delší příjmení než muži. Nulová hypotéza říká, že pohlaví a délka příjmení spolu nesouvisí; štítky F/M jsou tedy při pevných délkách zaměnitelné. Jako statistiku vezmeme

$$T = \bar{X}_F - \bar{X}_M.$$

V pozorovaných datech vyšlo $T_{\text{obs}} = 1.51$ znaku. Když náhodně permutujeme štítky F/M a vždy znovu spočítáme T, dostaneme permutační rozdělení pod H_0. V našich datech ani v jedné ze 100000 permutací nevyšla hodnota tak velká, tedy $p < 10^{-5}$.



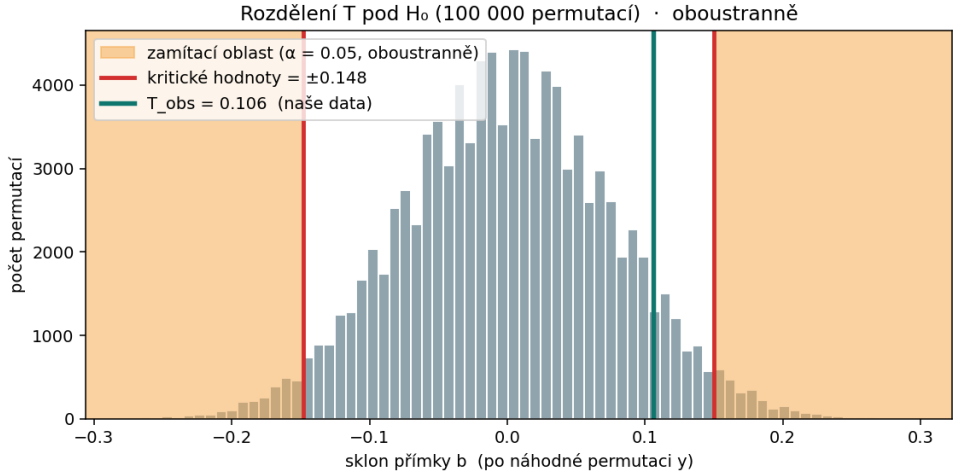
Obrázek 11: Permutační rozdělení rozdílu průměrných délek příjmení při náhodném přiřazování štítků F/M.

Příklad: délka jména a příjmení.

Jiná otázka je, zda délka křestního jména souvisí s délkou příjmení. Jako testovou statistiku můžeme vzít sklon regresní přímky. Máme-li dvojice (x_i, y_i), pod nulovou hypotézou žádné asociace držíme x_i fixní a uvažujeme permutace

$$(x_i, y_{\pi(i)}), \qquad \pi \in S_n.$$

Pro každou permutaci spočítáme statistiku T_π, například sklon regresní přímky, a pozorovanou hodnotu porovnáme s tímto permutačním rozdělením. V tomto příkladu vyšlo $p \approx 0.16$, takže nulovou hypotézu nezamítáme.



Obrázek 12: Permutační test sklonu: podíl permutací se stejně extrémním sklonem dává p-hodnotu přibližně 0.16.

Logika je stejná jako u z-testu:

1. zformulujeme H_0,
2. zvolíme statistiku T,
3. odvodíme nebo nasimulujeme rozdělení T za H_0,
4. podíváme se, zda pozorovaná hodnota leží v extrémní části tohoto rozdělení.

Rozdíl je jen v tom, že nulové rozdělení nevzniká z normální aproximace, ale z permutací, které jsou podle nulové hypotézy zaměnitelné.

2.6 Síla testu a experimentální design

Síla testu (anglicky *power*) je pravděpodobnost, že test zamítne H_0, když ve skutečnosti platí konkrétní alternativa. Nejčistší smysl má při porovnávání testů proti pevně určené alternativě:

$$\pi(\theta) = \mathbb{P}_\theta(\text{zamítneme } H_0), \qquad \theta \in \Theta_1.$$

Hladina α kontroluje falešné pozitivní závěry. Síla řeší opačnou otázku: jak často test skutečný signál zachytí. Chyba druhého druhu při konkrétní alternativě je $1 - \pi(\theta)$.

2.6.1 Porovnání testů na příkladu s mincí

Testujme

$$H_0 : p = 0.5 \qquad \text{proti} \qquad H_1 : p > 0.5$$

na základě 100 hodů mincí. Při alternativě $p = 0.6$ dostaneme například:

pravidlo	hladina	síla při $p = 0.6$	komentář
#hlav ≥ 59 na 100 hodech	≈ 0.044	≈ 0.623	rozumný test
#hlav ≥ 32 jen na prvních 50 hodech	≈ 0.032	≈ 0.336	zahodili jsme polovinu dat
zamítej náhodně s pravděpodobností 0.05	0.05	0.05	správná hladina, nulová citlivost

Tři testy mohou mít podobnou hladinu, ale velmi odlišnou schopnost zachytit skutečný signál. Hladina sama tedy neříká, zda je test dobrý.

2.6.2 Volba testu a návrhu experimentu

Sílu testu ovlivňuje velikost efektu, velikost vzorku, šum v měření a také to, zda test odpovídá mechanismu vzniku dat. U průměrů a podílů se pořád vrací měřítko $1/\sqrt{n}$: čtyřikrát větší vzorek dává zhruba dvakrát menší směrodatnou chybu.

situace	rozumný test	poznámka
Průměr, známá σ	z-test	silný, ale spoléhá na známý rozptyl a normální model
Průměr, neznámá σ	t-test	přirozená náhrada z-testu; kvůli odhadu rozptylu má těžší ocasy
Dvě randomizované skupiny	permutační nebo dvouvýběrový t-test	permutační test využívá zaměnitelnost; t-test přidává modelové předpoklady
Bez kontrolní skupiny	žádný test nevyřeší designový problém	zlepšení může být placebo, časový trend nebo regrese k průměru
S kontrolní skupinou	test rozdílu mezi skupinami	testujeme efekt zásahu proti přirozenému vývoji kontrol

Tabulka 1: Síla testu není jen vlastnost vzorce; často rozhoduje kvalita návrhu experimentu.

2.6.3 Placebo, regrese k průměru a kontrolní skupina

V reálných experimentech nestačí dívat se jen na změnu v jedné skupině. Zlepšení může vzniknout i bez účinného zásahu:

- **regrese k průměru**: když vybereme extrémní případy, při dalším měření často vypadají méně extrémně už jen z náhody,
- **placebo efekt**: lidé se často cítí lépe i bez účinné látky, protože očekávají zlepšení,
- **změna chování při měření**: samotná účast v experimentu může změnit to, co lidé dělají nebo reportují,
- **časový trend**: situace se mohla změnit z důvodů nesouvisejících se zásahem.

Název *regrese k průměru* je historický: Galton pozoroval, že extrémně vysocí rodiče mívají děti také vysoké, ale v průměru méně extrémní. Slovo "regrese" zde tedy neznamená fitování přímky jako v lineární regresi, ale návrat extrémních pozorování blíže k populačnímu průměru při opakovaném měření.

U léků na depresi je to zásadní problém: lidé často vyhledají pomoc ve chvíli, kdy je jim mimořádně špatně, a část zlepšení by nastala i bez nové léčby. Podobně když umístíme výstražné značky na místa, kde bylo v posledních letech neobvykle mnoho nehod, počet nehod tam může později klesnout i bez účinku značky, protože extrémní období se samo vrací blíže běžnému průměru.

Proto v medicíně a sociálních vědách potřebujeme *kontrolní skupinu*. Ideální je *randomizovaná kontrolovaná studie* (anglicky *randomized controlled trial*, zkratka RCT):

1. vhodné jednotky náhodně rozdělíme do skupiny se zásahem a kontrolní skupiny,
2. skupina se zásahem dostane sledovaný zásah,
3. kontrolní skupina dostane placebo, standardní péči nebo žádný zásah podle otázky,
4. porovnáváme rozdíl mezi skupinami, ne jen změnu uvnitř jedné skupiny.

Randomizace pomáhá vyrovnat známé i neznámé rušivé faktory mezi skupinami.

Kontrolní skupiny jsou užitečné i mimo medicínu. V průzkumech o konspiračních teoriích se například ukázalo, že část lidí souhlasí i s úplně vymyšleným tvrzením: 30.6% respondentů připustilo,

že George W. Bush mohl mít podíl na fiktivní havárii v North Dakotě. Takový kontrolní bod pomáhá odlišit specifickou víru v konkrétní teorii od obecné tendence souhlasit s podezřelými tvrzeními.

Právě pro porovnávání skupin existují standardní testy: varianty z-testu a t-testu pro rozdíly průměrů nebo podílů, případně permutační testy. Pokud máme randomizované přiřazení do treatment a control skupiny, permutační test je obzvlášť přirozený: pod nulovou hypotézou žádného efektu jsou štítky skupin zaměnitelné.

2.6.4 Simpsonův paradox

Agregovaná data mohou tvrdit opak toho, co platí uvnitř každé podskupiny zvlášť. To je jádro Simpsonova paradoxu.

Může se například stát, že průměrná životní spokojenost bělochů se za posledních deset let zvýšila a průměrná životní spokojenost nebělochů se také zvýšila, ale celkový průměr v populaci přesto klesl, protože se změnilo složení populace. Podobně může nový lék v agregovaných datech vypadat lépe než starý, i když je ve skutečnosti horší v každé věkové skupině, pokud byl testován hlavně na mladších pacientech s lepší prognózou.

Toto je složitější verze problému kontrol. V praxi často nemáme dokonale randomizovanou kontrolní a experimentální skupinu. Spíše porovnáváme dva průzkumy nebo dva experimenty, které vznikly na jiných profilech lidí. Telefonický a internetový volební průzkum mohou zasáhnout jiné věkové a vzdělanostní skupiny. Dva léky mohou být testované v různých nemocnicích; do jedné nemocnice třeba častěji přicházejí kritické případy. Pak může horší lék vypadat v agregátu lépe jen proto, že byl testován na zdravější populaci.

Simpsonův paradox není matematická chyba. Je to varování, že otázka "máme agregovat?" je věcná otázka o mechanismu vzniku dat. Někdy nás zajímá kontrolované srovnání uvnitř podskupin, typicky když porovnáváme účinnost léků. Jindy je relevantní právě agregát: pokud je celá populace chudší, je to důležitý ekonomický ukazatel i v situaci, kdy se každé podskupině zvlášť vede lépe. Správná úroveň agregace závisí na tom, co chceme odhadnout.

Praktická pointa

Dobrá statistika nekončí u algebraického výpočtu. Stejně důležitý je návrh experimentu, kontrolní skupina, randomizace, měření správné veličiny a kontrola toho, zda agregace neschovává důležitou strukturu dat.

3 Intervalové odhady

3.1 Definice konfidenční množiny

Bodový odhad dává jedno číslo. Často ale chceme vyjádřit i nejistotu. Místo jedné hodnoty proto vrátíme množinu hodnot parametru, které jsou s daty ještě kompatibilní.

Smysl konstrukce je dlouhodobý: když budeme stejný postup používat opakovaně, správná hodnota parametru má typicky ležet uvnitř vyrobeného intervalu. Konkrétní interval po pozorování dat už je pevný; buď parametr obsahuje, nebo ne. Garance říká, že metoda, která intervaly vyrábí, se trefuje dost často.

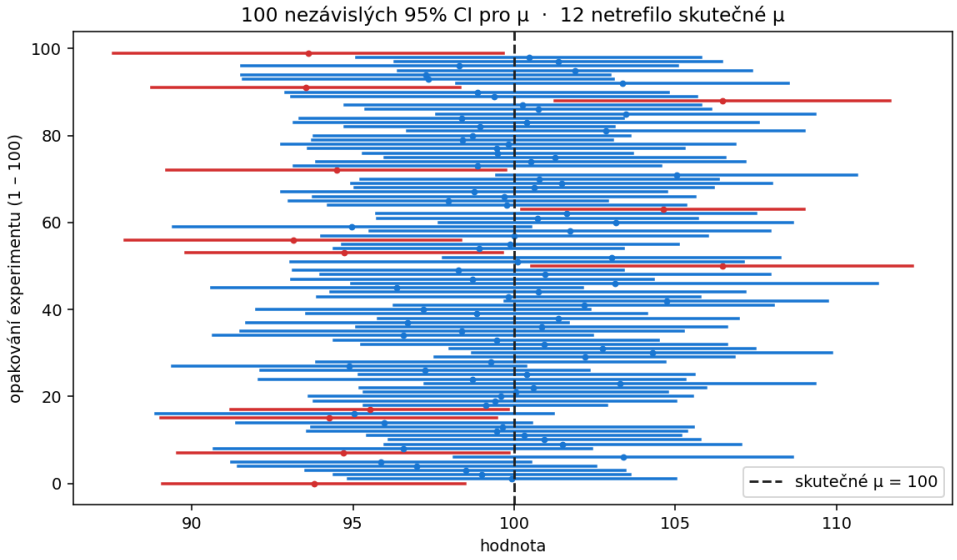
Definice. *Intervalový odhad* nebo obecněji *konfidenční množina* (anglicky *confidence set*) parametru θ na hladině α je pravidlo, které každému vzorku přiřadí množinu

$$C(X_1, \dots, X_n) \subseteq \Theta$$

tak, že pro každou skutečnou hodnotu $\theta \in \Theta$ platí

$$\mathbb{P}_\theta(\theta \in C(X_1, \dots, X_n)) \geq 1 - \alpha.$$

Číslo $1 - \alpha$ se nazývá pravděpodobnost pokrytí (anglicky *coverage probability*) nebo konfidenční hladina (anglicky *confidence level*).



Obrázek 13: Konfidenční hladina popisuje dlouhodobé pokrytí: při opakování experimentu většina intervalů obsahuje skutečnou hodnotu parametru.

Když je $C(X_1, \ldots, X_n)$ interval, píšeme často

$$C(X_1, \ldots, X_n) = [L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)].$$

Pak požadavek pokrytí zní

$$\mathbb{P}_\theta(L(X) \leq \theta \leq U(X)) \geq 1 - \alpha.$$

Interpretace

Náhodný je interval, ne parametr. Parametr je fixní, ale při opakování experimentu by se měnila data a tím i výsledný interval.

3.2 Jak číst 95% interval

Formulace

$$\mathbb{P}_\theta(\theta \in C(X_1, \ldots, X_n)) \geq 0.95$$

neznamená, že po pozorování konkrétních dat je pravděpodobnost pravdivosti výroku "θ leží v našem intervalu" rovna 95%. V klasickém, frekventistickém čtení je θ fixní a náhodný je postup, který interval vyrábí.

Správná interpretace je:

> Kdybychom celý experiment mnohokrát opakovali a pokaždé znovu použili stejný konstrukční postup, alespoň 95% takto vzniklých intervalů by obsahovalo skutečný parametr.

Po pozorování konkrétních dat už konkrétní interval buď skutečný parametr obsahuje, nebo ne. Nevíme kterou možnost, ale klasická pravděpodobnostní garance se týká dlouhodobého chování postupu.

Typicky budeme používat oboustranné intervaly $[L(X), U(X)]$. Existují ale i jednostranné konfidenční meze, například $(-\infty, U(X)]$ nebo $[L(X), \infty)$. Volba odpovídá tomu, zda by odpovídající test byl oboustranný nebo jednostranný.

Teď si ukážeme konstrukci intervalů z testů. Myšlenka je jednoduchá: pro každou hypotetickou hodnotu parametru uděláme test a v intervalu necháme právě ty hodnoty, které test nezamítne. Po konkrétním příkladu si obecně dokážeme, proč tento postup dává správné pokrytí.

3.3 Příklad: intervaly pro průměr

Nechť

$$X_1, \dots, X_n$$

jsou nezávislé a každá veličina má rozdělení $N(\mu, \sigma^2)$, kde σ je známé. Pro každou hypotetickou hodnotu μ_0 testujeme

$$H_0 : \mu = \mu_0$$

oboustranným z-testem. Testová statistika je

$$Z(\mu_0) = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Na hladině α test nezamítá právě tehdy, když

$$|Z(\mu_0)| \le z_{1-\alpha/2}.$$

Dosadíme a upravíme nerovnost:

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| \le z_{1-\alpha/2} \quad \Longleftrightarrow \quad \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Množina všech hodnot μ_0, které z-test nezamítne, je tedy

$$C(X) = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

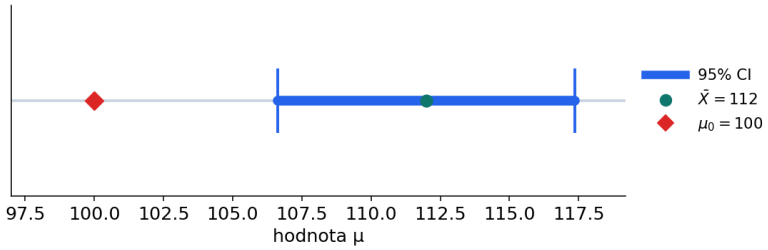
konfidenční interval pro μ s pokrytím $1 - \alpha$. Interval jsme tedy nezískali jako nový objekt, ale jako inverzi rodiny z-testů.

Pro $\alpha = 0.05$ je $z_{0.975} \approx 1.96$, takže

$$\mathrm{CI}_{95\%}(\mu) = \left[\bar{X} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{X} + 1.96\frac{\sigma}{\sqrt{n}} \right].$$

Při $\bar{X} = 112$, $\sigma = 15$, $n = 30$ vyjde

$$\mathrm{CI}_{95\%}(\mu) \approx [106.63,\ 117.37].$$



Obrázek 14: Ekvivalentní formulace téhož výsledku: buď test zamítne $H_0 : \mu = 100$, nebo řekneme, že 100 neleží v 95% intervalu.

Neznámý rozptyl: t-interval.

Když σ neznáme, použijeme

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

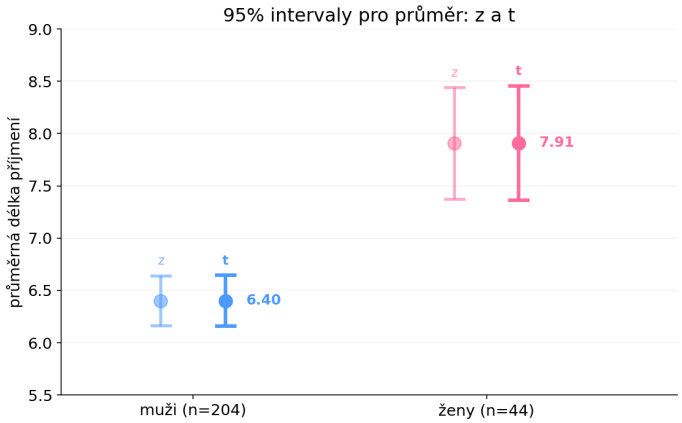
Za normálního modelu platí

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

Stejnou kvantilovou úpravou dostaneme

$$\mathrm{CI}_{1-\alpha}(\mu) = \left[\bar{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}},\ \bar{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right].$$

Pro malé vzorky je tento interval širší než z-interval, protože t-rozdělení má těžší chvosty a navíc odhadujeme rozptyl.



Obrázek 15: Na stejných datech můžeme dostat mírně odlišné intervaly podle zvolené konstrukce.

3.3.1 Nezamítací oblast

Pro test hypotézy

$$H_0 : \theta = \theta_0$$

na hladině α označme nezamítací oblast

$$A(\theta_0) = \{x : \text{test při datech } x \text{ nezamítne } H_0 : \theta = \theta_0\}.$$

Protože test má hladinu α, při pravdivé hypotéze $\theta = \theta_0$ platí

$$\mathbb{P}_{\theta_0}(X \notin A(\theta_0)) \leq \alpha.$$

Ekvivalentně

$$\mathbb{P}_{\theta_0}(X \in A(\theta_0)) \geq 1 - \alpha.$$

Nezamítací oblast je množina v prostoru dat. Inverze testu ji převrátí na množinu v prostoru parametrů.

3.3.2 Z testů na intervaly

Teď dokážeme obecný princip použitý v předchozím příkladu: kdykoliv máme pro každou hypotetickou hodnotu parametru test na hladině α, můžeme z těchto testů udělat konfidenční množinu. Do intervalu patří právě ty hodnoty parametru, které by odpovídající test při pozorovaných datech nezamítl.

Věta: interval jako inverze testů

Nechť pro každé $\theta_0 \in \Theta$ máme test hypotézy

$$H_0 : \theta = \theta_0$$

na hladině α s nezamítací oblastí $A(\theta_0)$. Definujme

$$C(X) = \{\theta_0 \in \Theta : X \in A(\theta_0)\}.$$

Pak $C(X)$ je konfidenční množina pro θ s pokrytím alespoň $1 - \alpha$.

Důkaz. Fixujme skutečnou hodnotu parametru θ. Z definice $C(X)$ platí ekvivalence

$$\theta \in C(X) \quad \Longleftrightarrow \quad X \in A(\theta).$$

Pravděpodobnost obou stran při skutečném parametru θ je tedy stejná:

$$\mathbb{P}_\theta(\theta \in C(X)) = \mathbb{P}_\theta(X \in A(\theta)).$$

Protože $A(\theta)$ je nezamítací oblast testu hypotézy $H_0 : \theta = \theta$ na hladině α, máme

$$\mathbb{P}_\theta(X \in A(\theta)) \geq 1 - \alpha.$$

Tedy

$$\mathbb{P}_\theta(\theta \in C(X)) \geq 1 - \alpha.$$

To je přesně definiční podmínka konfidenční množiny. □

3.3.3 Z intervalů na testy

Platí i opačný směr: každý konfidenční set definuje rodinu testů.

Věta: test jako inverze intervalu

Nechť $C(X)$ je konfidenční množina pro θ s pokrytím alespoň $1 - \alpha$. Pro každé $\theta_0 \in \Theta$ definujme test hypotézy

$$H_0 : \theta = \theta_0$$

tak, že zamítne právě tehdy, když

$$\theta_0 \notin C(X).$$

Pak má tento test hladinu nejvýše α.

Důkaz. Při pravdivé hypotéze $\theta = \theta_0$ je pravděpodobnost zamítnutí

$$\mathbb{P}_{\theta_0}(\theta_0 \notin C(X)).$$

Protože $C(X)$ má pokrytí alespoň $1 - \alpha$, platí

$$\mathbb{P}_{\theta_0}(\theta_0 \in C(X)) \geq 1 - \alpha.$$

Po doplnění do jedničky dostaneme

$$\mathbb{P}_{\theta_0}(\theta_0 \notin C(X)) \leq \alpha.$$

Test tedy má hladinu nejvýše α. □

Hlavní souvislost

Hodnota parametru leží v $100(1-\alpha)\%$ konfidenčním intervalu právě tehdy, když by odpovídající test hypotézy $H_0 : \theta = \theta_0$ na hladině α tuto hodnotu nezamítl.

3.4 Bootstrap

U průměru, podílu nebo normální lineární regrese umíme často rozdělení estimátoru spočítat ručně. U složitějších estimátorů, například mediánu, kvantilu, korelace, sklonu po robustní regresi nebo parametru složitého modelu, může být přesné odvození nepraktické. Bootstrap je obecná simulační metoda, jak nejistotu odhadnout přímo z dat.

Ideální postup pro odhad nejistoty by vypadal takto:

1. známe skutečné populační rozdělení P,

2. z P vygenerujeme mnoho nových datasetů velikosti n,
3. na každém datasetu spočítáme stejný estimator $\hat{\theta}$,
4. z rozdělení těchto hodnot popíšeme směrodatnou chybu nebo interval.

Problém je, že skutečné P neznáme. Bootstrap ho nahradí empirickým rozdělením pozorovaných dat.

3.4.1 Empirické rozdělení

Po pozorování dat $x_1, \ldots, x_n$ definujeme empirické rozdělení

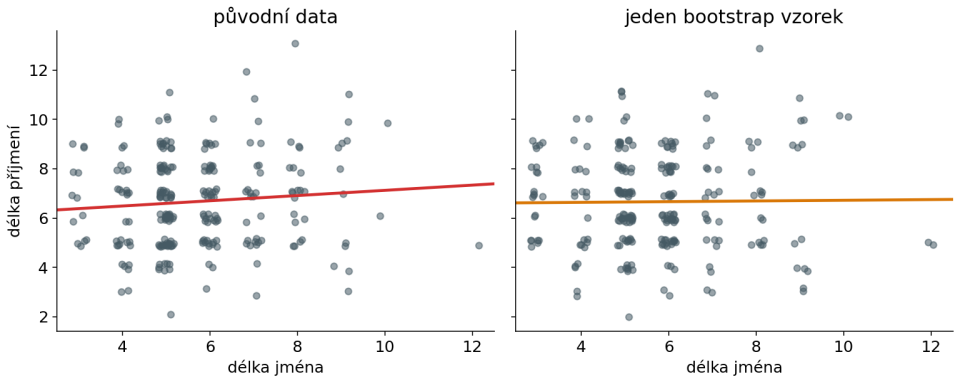
$$\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

kde δ_{x_i} je bodová pravděpodobnost v pozorování x_i. Jinými slovy: empirické rozdělení dává každému pozorovanému bodu pravděpodobnost $1/n$.

Bootstrapový vzorek

$$X_1^*, \ldots, X_n^*$$

je náhodný výběr velikosti n z empirického rozdělení $\hat{P}_n$, tedy výběr *s vracením* z původních dat. Některé původní body se v bootstrapovém vzorku objeví vícekrát a některé vůbec.



Obrázek 16: Bootstrapový vzorek vzniká losováním s vracením z původních pozorování.

3.4.2 Základní neparametrický bootstrap

Nechť estimator zapisujeme jako

$$\hat{\theta} = s(X_1, \ldots, X_n),$$

kde s je nějaká funkce dat. Bootstrapový postup:

1. Vylosuj s vracením bootstrapový vzorek $X_1^*, \ldots, X_n^*$ z původních dat.
2. Spočítej bootstrapovou verzi estimatoru

$$\hat{\theta}^* = s(X_1^*, \ldots, X_n^*).$$

3. Opakuj kroky 1 a 2 mnohokrát, například $B = 1000$ nebo $B = 10000$ krát.
4. Rozdělení hodnot $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$ použij jako aproximaci výběrového rozdělení $\hat{\theta}$.

Co bootstrap dělá

Bootstrap aproximuje rozdělení estimátoru tím, že neznámé populační rozdělení nahradí empirickým rozdělením pozorovaných dat.

3.4.3 Bootstrapová směrodatná chyba

Jakmile máme bootstrapové rozdělení, můžeme z něj odhadovat veličiny, které by jinak vyžadovaly znalost výběrového rozdělení estimatoru. U bodových odhadů jsme viděli, že variance estimatoru je důležitá, ale často ji neumíme spočítat ručně. Bootstrap ji odhaduje přímo z opakovaných bootstrapových hodnot

$$\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B},$$

jejichž průměr je

$$\bar{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{*b}.$$

Bootstrapový odhad směrodatné chyby je jejich výběrová směrodatná odchylka:

$$\widehat{\mathrm{se}}_{\mathrm{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^{*b} - \bar{\theta}^*)^2}.$$

Stejné bootstrapové rozdělení můžeme použít i ke konstrukci konfidenčních intervalů.

3.4.4 Percentilový bootstrapový interval

Nejjednodušší bootstrapový interval vezme percentily bootstrapového rozdělení. Pro 95% interval použijeme

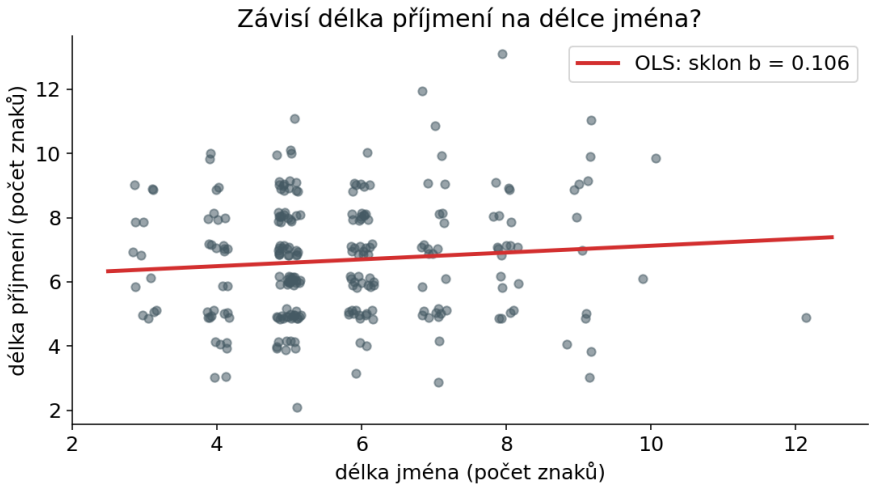
$$[q^*_{0.025},\ q^*_{0.975}],$$

kde q^*_γ je γ-kvantil hodnot $\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}$.

Tento interval je snadno pochopitelný a často prakticky užitečný. Není ale univerzálně nejlepší: u výrazně vychýlených nebo asymetrických bootstrapových rozdělení se používají i varianty jako basic bootstrap, bootstrap-t nebo BCa interval.

3.4.5 Příklad: sklon v regresi

V příkladu se ptáme, zda délka příjmení souvisí s délkou křestního jména. Odhad, který nás zajímá, je sklon regresní přímky b.



Obrázek 17: Estimator už není průměr, ale sklon regresní přímky.

Protože každé pozorování je dvojice

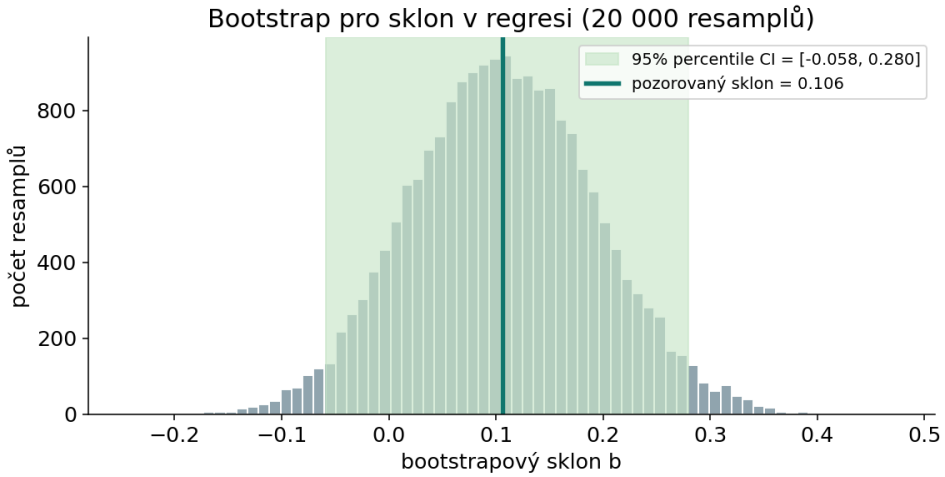
$$(x_i, y_i),$$

bootstrapujeme celé dvojice, ne zvlášť x a zvlášť y. V každém bootstrapovém vzorku znovu spočítáme sklon $\hat{b}^*$. Tím dostaneme empirické rozdělení odhadu sklonu.

Pro tato data vyšel percentile bootstrap interval přibližně

$$[-0.058,\ 0.280].$$

Nula leží uvnitř intervalu. To je konzistentní s tím, že permutační test nenašel silný signál proti nulové hypotéze nulového vztahu.



Obrázek 18: Bootstrapové rozdělení odhadu sklonu a z něj odvozený 95% interval.

3.4.6 Kdy bootstrap funguje dobře

Bootstrap obvykle funguje dobře, když:

- data jsou přibližně nezávislá a stejně rozdělená,
- vzorek rozumně reprezentuje populaci,
- estimator je dostatečně stabilní,
- nejde o extrémní extrapolaci do části rozdělení, kterou jsme skoro nepozorovali.

U závislých dat, například časových řad, obyčejný bootstrap rozbije strukturu závislosti. Tam se používají jiné varianty, například block bootstrap. U velmi malých vzorků může empirické rozdělení reprezentovat populaci špatně. U extrémních kvantilů bootstrap často podceňuje nejistotu, protože v datech chybí informace o nepozorovaných chvostech.

Teoretické garance bootstrapu jsou obecně těžší než samotný algoritmus. V jednoduchých hladkých situacích funguje velmi dobře, ale u hraničních, nespojitých nebo silně závislých problémů je potřeba ověřit předpoklady konkrétní bootstrapové varianty.