Pravděpodobnost vs. statistika

Teorie pravděpodobnosti

- pracuje s jednou nebo více teoretickými náhodnými veličinami, jejichž rozdělení známe
- odvozovali jsme charakteristiky těchto rozdělení atd.

Statistika

- pracuje s pozorováními (daty) $\leftrightsquigarrow$ realizace nějaké náhodné veličiny z neznámého rozdělení (tzv. náhodný výběr)
- pomocí dat chceme odhadovat např. střední hodnotu, rozptyl, hustotu atd.

Budeme zkoumat chování tzv. náhodného výběru.

Náhodný výběr

Definice

Nechť $X_1, \ldots, X_n$ jsou nezávislé a stejně rozdělené náhodné veličiny s distribuční funkcí F. Pak říkáme, že se jedná o *náhodný výběr* z rozdělení F.

- Náhodný výběr jsou měření téže charakteristiky prováděná na různých subjektech.
- Velký význam ve statistice (odhady, testy atd.).
- Bude nás zajímat chování a vlastnosti náhodného výběru.

Příklad

Příklad: Chceme zjistit znečištění řeky, a to pomocí měření koncentrace škodlivin (např. mědi) u zde žijících ryb.

- Logicky nemůžeme pochytat a proměřit úplně všechny ryby.
- Koncentraci škodlivin považujeme za náhodnou veličinu s nějakým neznámým rozdělením se střední hodnou μ.
- Vylovíme náhodně n ryb: naměřená koncentrace u těchto n ryb bude tvořit náhodný výběr.

Chtěli bychom odhadnout např. $\mu \rightsquigarrow$ Co je přirozené vzít jako odhad? Jaké má tento odhad vlastnosti?

Výběrový průměr

Definice

Výběrovým průměrem náhodného výběru $X_1, \ldots, X_n$ rozumíme náhodnou veličinu

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- výběrový průměr je náhodná veličina (kdybychom získali znovu jiný náhodný výběr, dostali bychom jiné hodnoty X_i a tudíž jiný výběrový průměr)
- lze tedy uvažovat jeho rozdělení, střední hodnotu, rozptyl a všechny ostatní charakteristiky

Vlastnosti výběrového průměru

Označme μ střední hodnotu a σ^2 rozptyl veličin $X_1, \ldots, X_n$

Věta

Platí:

1. $\mathrm{E}\overline{X}_n = \mathrm{E}X_i = \mu$.
2. $\mathrm{var}\,\overline{X}_n = \frac{1}{n}\mathrm{var}\,X_i = \frac{\sigma^2}{n}$.
3. Pochází-li náhodný výběr z normálního rozdělení $\mathrm{N}(\mu, \sigma^2)$, pak výběrový průměr má také normální rozdělení,

$$\overline{X}_n \sim \mathrm{N}(\mu_X, \frac{\sigma_X^2}{n}).$$

Normovaný průměr $\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma}$ má potom $\mathrm{N}(0, 1)$.

Příklad (znečištění řeky)

Předchozí věta nám (mimo jiné) říká:

- budeme-li opakovaně provádět experiment a v každém opakování změříme n ryb, pak v průměru bychom měli dostat skutečnou střední hodnotu μ
- variabilita průměru klesá se zvyšujícím se n
- čím více ryb proměříme, tím je menší variabilita průměru a tedy dostáváme přesnější hodnoty blíže hledanému μ

Vlastnosti výběrového průměru

Důkaz:

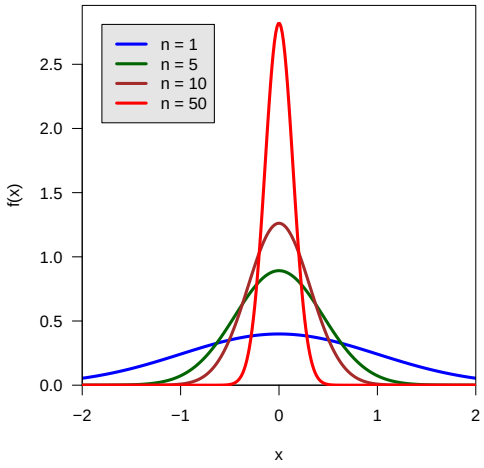
Dokážeme si pouze body 1 a 2.

1. $$\mathrm{E}\overline{X}_n = \mathrm{E}\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}X_i = \frac{n\mu}{n} = \mu.$$

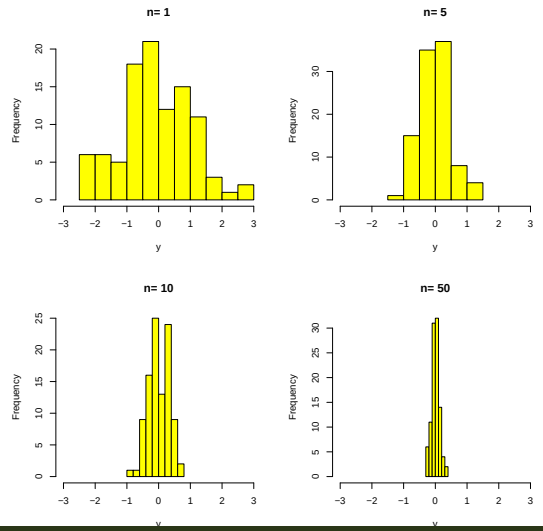
2. $$\mathrm{var}\,\overline{X}_n = \mathrm{var}\,\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n^2}\mathrm{var}\,\sum_{i=1}^{n} X_i =$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}\,X_i = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

(Využili jsme nezávislosti $X_1, \ldots, X_n$.)

Ilustrace vlastností výběrového průměru



Ilustrace vlastností výběrového průměru



Zákon velkých čísel

- Viděli jsme, že $\mathrm{E}\overline{X}_n = \mu$ a $\mathrm{var}\,\overline{X}_n = \sigma^2/n$. Je-li n hodně velké, pak je rozptyl $\overline{X}_n$ hodně malý, tj. hodnoty $\overline{X}_n$ kolísají jen velmi málo kolem střední hodnoty μ.
- Lze tedy očekávat, že pro nekonečně mnoho pozorování by průměr mohl být přímo roven μ.

Věta (Slabý zákon velkých čísel)

Mějme dán (nekonečný) náhodný výběr $X_1, X_2, \ldots$ z rozdělení se střední hodnotou $\mu < \infty$. Potom platí, že výběrový průměr $\overline{X}_n$ spočítaný z prvních n pozorování se s $n \to \infty$ přibližuje ke střední hodnotě μ ve smyslu

$$\lim_{n\to\infty} P[|\overline{X}_n - \mu| > \varepsilon] = 0 \text{ pro každé } \varepsilon > 0.$$

Význam zákona velkých čísel

- Spočítáme-li výběrový průměr z nekonečného náhodného výběru, dostaneme střední hodnotu μ
- Spočítáme-li výběrový průměr z konečného ale velkého náhodného výběru, nedostaneme přesně střední hodnotu, ale dostaneme číslo, které je střední hodnotě blízko.

Zákon velkých čísel (ZVČ)

- ukazuje, že střední hodnota je výsledek, který bychom dostali v průměru při nekonečném množství opakování pokusu
- proto je výběrový průměr opravdu „dobrý“ odhad střední hodnoty
- patří mezi tzv. limitní věty
- lze jej aplikovat všude, kde se vyskytuje výběrový průměr nějakých veličin (tj. např. na $1/n \cdot \sum_{i=1}^{n} X_i^2$ apod.)

Centrální limitní věta

- pro výběr z normálního rozdělení má $\overline{X}_n$ opět normální rozdělení
- v jiných případech bývá obtížné určit rozdělení $\overline{X}_n$
- stačilo by nám znát toto rozdělení alespoň přibližně

Věta (Centrální limitní věta)

Mějme dán (nekonečný) náhodný výběr $X_1, X_2, \ldots$ z rozdělení se střední hodnotou $\mu < \infty$ a rozptylem $\sigma^2 > 0$. Potom má náhodná veličina $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ přibližně normované normální rozdělení $\mathrm{N}(0,1)$ ve smyslu

$$\lim_{n\to\infty} P\left[\sqrt{n}\frac{(\overline{X}_n - \mu)}{\sigma} \le x\right] = \Phi(x) \text{ pro každé } x \in \mathbb{R},$$

kde Φ je distribuční funkce rozdělení $\mathrm{N}(0,1)$.

Význam centrální limitní věty

Centrální limitní věta (CLV)

- výběrový průměr se při velkém rozsahu výběru chová jako normálně rozdělená náhodná veličina
- ekvivalentní zápisy tvrzení:

$$\sqrt{n}\,\frac{\overline{X}_n - \mu_X}{\sigma_X} \dot{\sim} \mathrm{N}(0,1)$$

$$\sqrt{n}(\overline{X}_n - \mu_X) \dot{\sim} \mathrm{N}(0,\sigma_X^2)$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - \mu_X) \dot{\sim} \mathrm{N}(0,\sigma_X^2)$$

$$\overline{X}_n \dot{\sim} \mathrm{N}(\mu_X, \frac{\sigma_X^2}{n})$$

$$\sum_{i=1}^{n} X_i \dot{\sim} \mathrm{N}(n\mu_X, n\sigma_X^2)$$

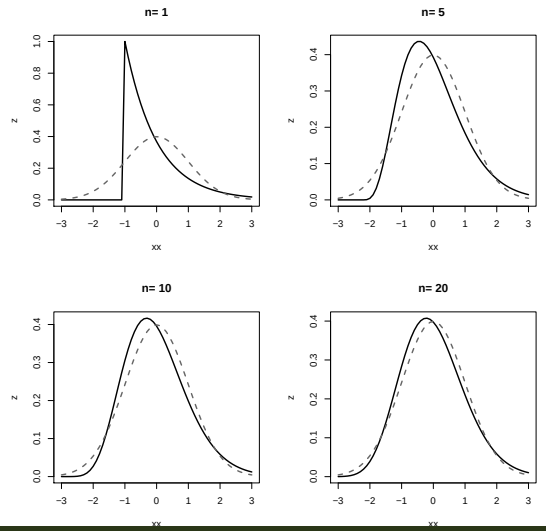
Význam centrální limitní věty

Centrální limitní věta (CLV)

- víme, že $\overline{X}_n$ má střední hodnotu μ a rozptyl σ^2/n
- veličina $\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}$ je normovaná tak, aby měla nulovou střední hodnotu a jednotkový rozptyl

- víme, že pokud X_i pocházejí z normálního rozdělení, pak $\overline{X}_n$ je také normální
- CLV: ať X_i pocházejí z jakéhokoli rozdělení ➡ výběrový průměr je při dostatečně velkém počtu pozorování vždy přibližně normální
- toho lze využít pro přibližné výpočty pravděpodobností, testování atd.

Hustota (normovaného) průměru z Exp(1)



Význam centrální limitní věty

Centrální limitní věta ukazuje, proč je normální rozdělení tak důležité

- řada věcí, s kterými budeme pracovat, má podle centrální limitní věty přibližně normální rozdělení
- řada veličin z praxe má rozdělení blízké normálnímu, neboť je lze vyjádřit nebo představit si jako součty či průměry velkého počtu nezávislých náhodných veličin