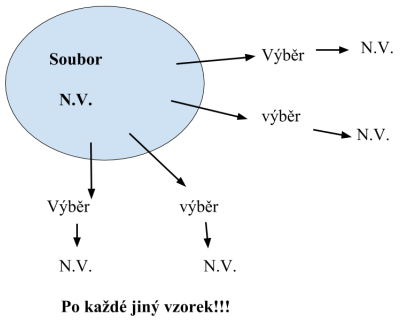
Přednáška 5 - soubor, výběr, bodový odhad, limitní věty

Musíme si uvědomit, že při pozorování náhodné veličiny nikdy nenaměříme úplně všechny její hodnoty. V každém případě to bude pouze nějaká část dat. Například, zajímá nás průměrný věk obyvatel ČR. Musíme každého obejít a zeptat se na věk. Než ale všech 10 milionů lidí obejdeme, tak někdo zestárne, umře, narodí se, což znamená, že musíme začít od začátku. Abychom takovému problému zabránili, musíme pořídit náhodný vzorek dat. Tomu vzorku dat budeme říkat výběr nebo náhodný výběr, a přesně toto bude tématem dnešní přednášky.

Výběr nebo náhodný výběr

Uvědomíme si, že po každém měření dostaneme odlišný vzorek dat. Například, měříme (ptáme se na věk) dneska na škole. Pak přijdeme měřit zítra – někdo už chybí a naopak přišel někdo jiný. To znamená, že hodnoty, které jsme naměřili jsou také náhodné veličiny.



Můžeme to ukázat na obrázku, kde máme soubor všech hodnot náhodné veličiny (věk obyvatel) a opakovaně provádíme její měření (pokaždé vyjde jiný vzorek dat). Z toho důvodu definice výběru zní takto:

Výběr je množina nezávislých a stejně rozdělených náhodných veličin. Označíme výběr písmenem X:

$$X = \underbrace{[X_1, X_2, \ldots, X_n]}_{\text{nezávislé+stejně rozdělené}}, \qquad \text{kde } n \text{ je počet hodnot výběru}$$

Proč potřebujeme stejně rozdělené náhodné veličiny?

> Měříme tu náhodnou veličinu, která nás v tuto chvíli zajímá. Například, pro zjištění věku obyvatel ČR nepůjdeme měřit do Polska.

Proč musí být nezávislé náhodné veličiny?

> Protože je potřeba počítat s tím, že výběr musí být reprezentativní. Není vhodné jít ptát se na věk obyvatel do mateřské školky, do domova pro seniory nebo do supermarketu dopoledne ve všední den. Výběr by v tom případě byl jednostranně zaměřen – dokážete si představit, jaký průměrný věk bychom takto spočítali. Z toho důvodu, musíme měřit, pokud možno, v nezávislé skupině obyvatel (např., MHD, obchodní centra o víkendu, zoologická zahrada, aquapark).

Statistika

Připomeneme, že cílem pořizování náhodného výběru je určení charakteristik souboru, tj pozorované náhodné veličiny. Například, předpokládáme, že věk obyvatel je náhodná veličina s normálním rozdělením

$$\text{věk} \sim N(\mu, \sigma^2),$$

kde střední hodnota μ a rozptyl σ^2 jsou parametry normálního rozdělení. To znamená, že pokud nás zajímá průměrný věk obyvatel, potřebujeme střední hodnotu μ. Kdybychom chtěli upřesnit, v jakém rozmezí od střední hodnoty se nachází naměřená data, potřebujeme rozptyl σ^2. Abychom tyto charakteristiky souboru mohli počítat (odhadnout) z výběru, zavedeme nový obecný pojem, kterému budeme říkat statistika.

Statistika je funkce výběru, hodnota které nám dá bodový odhad parametru (v našem případě buď μ nebo σ^2). Obecně označíme statistiku písmenem T, a to je tedy

$$T(X) = \hat{\theta},$$

kde θ je obecné označení parametrů, stříška ˆ znamená bodový odhad. Víme, že se výběr X při každém měření liší. To znamená, že statistika $T(X)$ nám nedá přesnou hodnotu parametru, ale při každém výpočtu dostaneme i trochu jiný bodový odhad $\hat{\theta}$.

Z toho vyplývá, že statistika T je také náhodná veličina se svým rozdělením $f(T)$.

Konstrukce statistiky je samostatná úloha, která se řeší například metodou momentů nebo metodou maximální věrohodnosti (pro doplnění teorie jsou tyto metody popsány ve skriptech na webu). Především se seznámíme s hotovými statistikami a jejich vlastnostmi.

Výběrový průměr

Výběrový průměr je průměr, který počítáme z výběru, což znamená, že to je funkce výběru a tím pádem to je jeden z příkladů statistiky. Víme, že průměr je

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

Víme, že výběr obsahuje náhodné veličiny. Z toho důvodu je výběrový průměr také náhodnou veličinou a při každém výpočtu se liší:

$$\underbrace{\bar{X}}_{N.V.} = \frac{1}{n}\sum_{i=1}^{n}\underbrace{X_i}_{N.V.}$$

Například, věk cestujících v metru je:

$$[12 \quad 36 \quad 4 \quad 41 \quad 15 \quad 16 \quad 28 \quad 35 \quad 33 \quad 61 \quad \ldots].$$

Pořídíme náhodným způsobem 3 výběry ze 3 hodnot a spočteme výběrový průměr. Vidíme, že v každém případě je jiný:

$$\begin{aligned}
\text{např., první výběr} \quad & X = [12 \quad 4 \quad 28], \quad \bar{X} = 14.67, \\
\text{druhý výběr} \quad & X = [41 \quad 35 \quad 33], \quad \bar{X} = 36.33, \\
\text{třetí výběr} \quad & X = [4 \quad 28 \quad 15], \quad \bar{X} = 15.67.
\end{aligned}$$

Tady je vidět, že výběrový průměr $\bar{X}$ je náhodná veličina. Budeme předpokládat, že $\bar{X}$ má normální rozdělení se střední hodnotou $E[\bar{X}]$ a rozptylem $D[\bar{X}]$.

Střední hodnota výběrového průměru

Střední hodnota výběrového průměru se rovná střední hodnotě souboru:

$$E[\bar{X}] = \mu.$$

Ukážeme si, proč je to tak:

Dáme $\bar{X}$ podle definice do závorek:

$$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] =$$

$\frac{1}{n}$ je konstanta, počítání se střední hodnotou je v tom případě $E[ax] = aE[x]$, tedy můžeme konstantu vyndat ze závorek:

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

dále střední hodnota součtu se rovná součtu středních hodnot $E[x+y] = E[x] + E[y]$, proto

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

$E[X_i]$ je střední hodnota souboru – naší náhodné veličiny (věku) podle definice, a je to μ (protože věk $\sim N(\mu, \sigma^2)$). Platí tedy

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{E\left[X_i\right]}_{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu.$$

Důležitá poznámka: Tato vlastnost výběrového průměru ukazuje, že má smysl pracovat s výběrem, protože střední hodnota jeho průměru je stejná jako průměr celého souboru. Pro náš příklad s věkem to znamená, že pokud budeme opakovaně měřit věk obyvatel ČR v MHD, naměříme sice v každém případě něco jiného, ale v průměru to bude odpovídat střední hodnotě obyvatel ČR, což je průměr ze všech dat.

Rozptyl výběrového průměru

Rozptyl výběrového průměru se rovná rozptylu souboru vydělenému počtem dat

$$D[\bar{X}] = \frac{\sigma^2}{n}.$$

Důkaz:

Dáme $\bar{X}$ podle definice do závorek:

$$D[\bar{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] =$$

$\frac{1}{n}$ je konstanta, počítání s rozptylem je v tom případě $D[ax] = a^2 D[x]$, tedy můžeme napsat:

$$= \frac{1}{n^2} D\left[\sum_{i=1}^{n} X_i\right] =$$

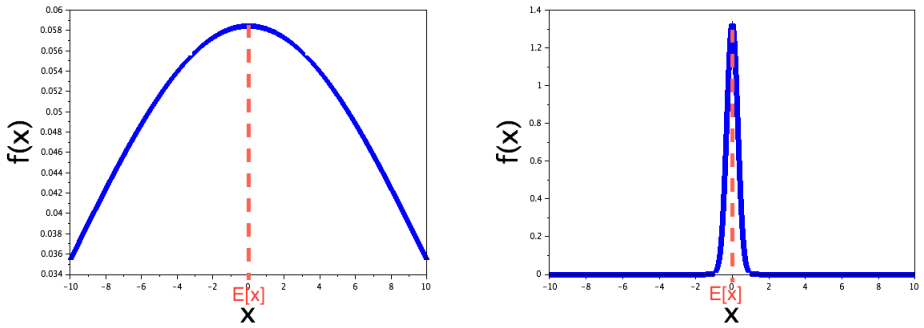
pro nezávislé náhodné veličiny platí $D[x+y] = D[x] + D[y]$, výběr je množina nezávislých veličin podle definice, proto

$$= \frac{1}{n^2}\sum_{i=1}^{n} D\left[X_i\right] =$$

$D[X_i]$ je rozptyl souboru, protože věk $\sim N(\mu, \sigma^2)$, a to je σ^2, tedy

$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{D\left[X_i\right]}_{\sigma^2} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Důležitá poznámka: Díky této vlastnosti v praxi se vyplatí používat co největší výběr. Čím víc máme dat (větší n), tím menší je rozptyl výběrového průměru, tím pádem je odhad přesnější a blíž ke střední hodnotě souboru.



Můžeme to ukázat na obrázku: vlevo je větší rozptyl, vpravo je malý rozptyl. Je vidět, že hodnoty vpravo budou blíž ke střední hodnotě.

Výběrový rozptyl

Další charakteristika, kterou dokážeme spočítat z výběru je výběrový rozptyl. Podobně jako výběrový průměr, to je náhodná veličina. Na rozdíl od rozptylu souboru σ^2, označíme výběrový rozptyl s^2 a počítáme ho následovně:

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

Všimněme si, že v porovnání s rozptylem, který jsme zavedli na první přednášce (tzv. souborový rozptyl), výběrový rozptyl se liší pouze $n-1$ ve jmenovateli. O tom jsme už mluvili i na cvičení. Tento rozdíl je nutný pro zachování vlastností rozptylu, o čemž si povíme později. Je nutné si ale poznamenat, že pro velký výběr je rozdíl v hodnotách rozptylů zanedbatelný: zkuste porovnat $\frac{1}{1000}$ a $\frac{1}{999}$.

Střední hodnota výběrového rozptylu

Střední hodnota výběrového rozptylu se rovná rozptylu souboru

$$E[s^2] = \sigma^2.$$

Pro náš příklad to znamená, že pokud si budeme dělat záznamy o věku obyvatel na základě výběrů z metra, průměrná rozptýlenost hodnot výběru bude stejná jako kdybychom obešli všechny obyvatele.

Výběrový podíl

Výběrový podíl je charakteristika výběru, se kterou můžeme pracovat v případě alternativního rozdělení souboru. Připomeneme si, že náhodná veličina, která má alternativní rozdělení, může nabývat dvou možných hodnot: 1 (úspěch) a 0 (neúspěch).

Výběrový podíl označíme p a spočteme ho takto:

$$p = \frac{n^+}{n},$$

kde n^+ je počet úspěchů ve výběru. Například, zajímá nás podíl maminek s kočárkem mezi cestujícími. V tomto případě, je maminka s kočárkem úspěch. Vydělíme počet maminek s kočárkem počtem dat výběru a dostaneme výběrový podíl. Všimněme si, že to je zároveň i pravděpodobnost úspěchu, tj., že náhodně vybraný cestující bude maminka s kočárkem.

Limitní věty

Zavedli jsme dva nové pojmy – soubor a výběr.

$$\mu, \sigma^2 \qquad \bar{X}, s^2$$

Ukážeme, jak spolu souvisí jejich charakteristiky, což jsme si už částečně poznamenali. Formálně vztah výběru a souboru popíšeme pomocí limitních vět.

Zákon velkých čísel

Při velkém rozsahu výběru $\underbrace{n \to \infty}_{\text{v praxi } n>30}$ se hodnoty výběrových charakteristik neomezeně blíží k charakteristikám souboru:

$$\bar{X} \to \mu$$
$$s^2 \to \sigma^2$$

Poznámka: Tato limitní věta říká, že pro určení charakteristik souboru (například, průměrného věku obyvatel) stačí pracovat s velkým náhodným výběrem z MHD. Pak bude výběrový průměr dobrým odhadem střední hodnoty souboru.

Centrální limitní věta

Při velkém rozsahu výběru $\underbrace{n \to \infty}_{\text{v praxi } n>30}$ má výběrový průměr přibližně normální rozdělení

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Poznámka: V praxi použití této limitní věty naznačuje tři důležité body:

- Pokud máme velký výběr z metra, můžeme předpokládat, že se výběrový průměr bude nacházet kolem své střední hodnoty, která se rovná střední hodnotě souboru μ (tj., je stejná jako pro všechny obyvatele).
- Čím větší je výběr ($\Uparrow n$), tím menší je rozptyl ($\Downarrow \frac{\sigma^2}{n}$), což znamená, že jsme zase blíž ke střední hodnotě souboru μ.
- I když hodnoty velkého výběru nepochází z normálního rozdělení, bude mít výběrový průměr $\bar{X}$ přibližně normální rozdělení.

Vlastnosti statistiky

Takovým způsobem jsme probrali některé existující statistiky (připomeneme, že statistika je funkce výběru). Abychom mohli pomocí statistiky dostat dobrý bodový odhad, musí mít statistika tři následující vlastnosti.

Nestrannost

Statistika T poskytuje nestranný bodový odhad parametru, jestliže se její střední hodnota rovná tomuto parametru:

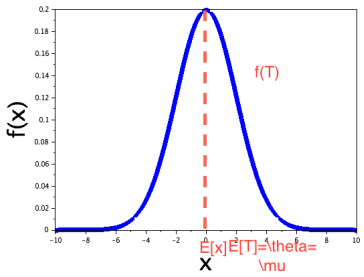
$$E[T] = \theta.$$

Například, výběrový průměr $\bar{X}$ je statistika a je to funkce výběru

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

Ukázali jsme, že $\underbrace{E[\bar{X}]}_{E[T]} = \underbrace{\mu}_{\theta}$ – to znamená, že $\bar{X}$ je nestranná statistika.

Tato vlastnost říká, že bodový odhad je v průměru přesný.



Například, výběrový rozptyl je nestranná statistika, protože

$$E[s^2] = \sigma^2.$$

Právě pro zachování nestrannosti je potřeba definovat výběrový rozptyl $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Dá se ukázat, že pro rozptyl definovaný jako $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ neplatí $E[s^2] = \sigma^2$ a rozptyl by nebyl nestranným odhadem.

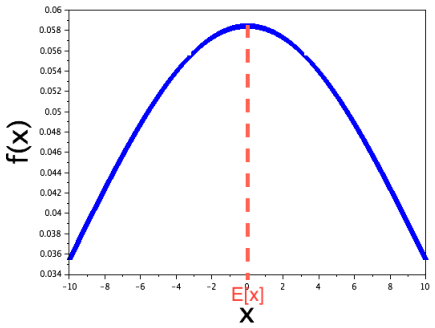
Konzistence

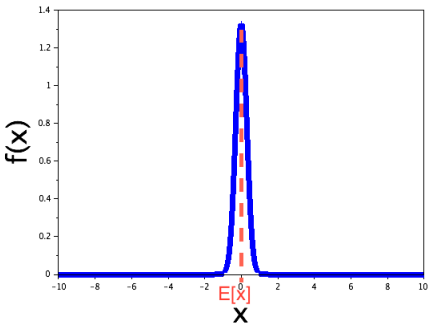
Statistika je konzistentní, pokud s rostoucím rozsahem výběru $n \to \infty$ se rozptyl statistiky blíží k nule $D[T] \to 0$.

Například, výběrový průměr $\bar{X}$ je statistika. Rozptyl výběrového průměru je

$$D[T] = D[\bar{X}] = \frac{\sigma^2}{n} \quad - \quad \text{to znamená, pro } \Uparrow n, \Downarrow \frac{\sigma^2}{n}, \text{ tedy to je konzistentní statistika.}$$

V praxi to znamená, čím více dat, tím je odhad přesnější. Například, na obrázku vlevo s velkým rozptylem pozorujeme data v rozmezí od -10 do 10 kolem nulové střední hodnoty. Na obrázku vpravo s menším rozptylem se data nachází na intervalu od -1 do 1 kolem 0.





Vydatnost

Tato vlastnost říká, že čím menší je rozptyl statistiky $D[T]$, tím je vydatnější, tj., přesnější:

$$D[T] \Downarrow, \text{ přesnost odhadu } \Uparrow$$

Například, v případě dvou nestranných statistik T_1 a T_2 volíme tu statistiku, která bude mít menší rozptyl a bude tedy vydatnější, tj.,

$$\text{pokud} \quad D[T_1] < D[T_2], \qquad T_1 \text{ – je vydatnější.}$$

Poznámka: pro doplnění teorie na webu jsou k dispozici skripta ze statistiky.