Charakteristiky náhodné veličiny

Příklad PASTELKY:

Jaká je průměrná délka pastelek ve vašem výběru?

Jaký je průměrný počet červených pastelek ve vašem výběru?

Jaké barvy a s jakou pravděpodobností se vyskytují ve vašem výběru?

Tato čísla jsou výběrové odhady. Čeho?

→ Populačního průměru. Statistický název je střední hodnota.

Podobně můžeme spočítat výběrový rozptyl a směrodatnou odchylku.

To, co umíme spočítat, jsou výběrové odhady.
Jak vypadají jejich teoretické protějšky – populační parametry?

Charakteristiky náhodné veličiny (populační charakteristiky)

Střední hodnota μ_X, EX [expected value, mean value]

$$\mu_X = EX = \sum_{j=1}^{\infty} x_j^* P(X = x_j^*)$$

veličina **X** s diskrétním rozdělením
… x_j^* je typická hodnota

$$\mu_X = EX = \int_{-\infty}^{\infty} x \cdot f(x) dx$$

veličina **X** se spojitým rozdělením
f(x) je hustota

- Je to „vážený průměr“ všech možných hodnot veličiny **X**, vážíme pomocí pravděpodobnosti, že hodnota nastane.
- Jiná interpretace: je to těžiště možných hodnot veličiny **X**.
- Odvodíme to cestou z výběrového průměru na populační průměr:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{\text{celá populace}} \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{j=1}^{m} n_j \cdot x_j^* = \sum_{j=1}^{m} \frac{n_j}{N} \cdot x_j^*$$

relativní četnost $x_j^* \approx$ pravděpodobnost

Populační medián [population median]

Stále označuje „prostřední“ hodnotu, ale populace může být až nekonečná, nejsme tedy schopni jedince očíslovat.

Stále platí, že 50 % všech hodnot je menších a 50 % větších než medián, ale definice je vedena přes pravděpodobnost:

$$P(X < \text{medián}) = 0.50 \quad a \quad P(X > \text{medián}) = 0.50$$

(Populační) rozptyl σ^2_X, *var X* [variance, dispersion]

- Různé způsoby zápisu:

$$\sigma^2_X = varX = D(X) = E(X - EX)^2 = E(X - \mu_X)^2$$
$$= \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f(x)dx$$

- Opět vážený čtverec vzdálenosti náhodné veličiny ***X*** od její střední hodnoty ***EX***, vážíme pravděpodobností, že hodnota X nastane**.**
- Rozptyl má jiné měřítko než původní náhodná veličina.

(Populační) směrodatná odchylka σ_X, *sd X* [standard deviation]

$$\sigma_X = \sqrt{\sigma^2_X} = SD(X) = \sqrt{D(X)}$$

- Odmocnina z populačního rozptylu
- Má stejné měřítko jako původní náhodná veličina

Míry diverzity, entropie

- Variabilita veličiny na nominální stupnici (např. barva květů)

Měli jsme entropii výběrového souboru:

$$H = -\sum_{j=1}^{m} \frac{n_j}{n} \cdot ln \frac{n_j}{n}$$

! Kategorie jsou očíslovány, ale výpočet charakteristik není ovlivněn pořadím kategorií.

Shannonova entropie je populační obdoba, místo relativních četností má přímo pravděpodobnosti

$$H = -\sum_{j=1}^{m} p_j \cdot ln(p_j)$$

Používá se jako míra diverzity, míra bohatosti společenstva.

Simpsonův index (Giniho index)

$$1 - \sum_{j=1}^{m} {p_j}^2 = \sum_{j=1}^{m} p_j \cdot (1 - p_j)$$

Pravá strana rovnice popisuje prst mezidruhového setkání při neomezeně velkém společenstvu.

Některé modely rozdělení pravděpodobností

Vizte samostatnou prezentaci „Některá rozdělení“.

Bodový odhad parametru [point estimate of the parameter]

Základní předpoklad dalšího odvozování:
mám výběr ***n*** hodnot $(X_1, X_2, X_3, \dots, X_n)$, které jsou ***iid.***, tedy vzájemně nezávislé a všechny pocházejí ze stejného rozdělení prstí.

K odhadu typické hodnoty (charakteristika polohy) nejčastěji používáme

výběrový průměr $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ [sample mean]

Protože výběrový průměr je náhodná veličina, má smysl se ptát:

a) jaká je jeho střední hodnota [expected value of the estimate]
b) jaký je jeho rozptyl [variance of the estimate]
c) jaká je jeho směrodatná odchylka [standard error of the estimate]

Populační charakteristiky průměru (odvození dále):

$$E\bar{X} = \mu \qquad var\,\bar{X} = \frac{\sigma^2}{n} \qquad sd\,\bar{X} = \frac{\sigma}{\sqrt{n}}$$

Odvození pro výběrový průměr:

(a) Střední hodnota výběrového průměru:

$$E\bar{X} = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n}\cdot n \cdot \mu = \mu$$

vlastnost střední hodnoty: $\mathbf{E}(X + Y) = EX + EY$

- tento odhad je nestranný, protože $E\bar{X} = \mu$

Odvození pro výběrový průměr:

(b) Rozptyl výběrového průměru:

$$var\bar{X} = var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\, var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} varX_i\right) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \frac{\sigma^2}{n}$$

$var(\beta \cdot X) = \beta^2 \cdot varX$

(1) všechna X_i jsou ***iid.***, proto $cov(X_i, X_j) = 0$ pro $\forall i, j$
(2) $var(X + Y) = varX + varY + 2cov(X, Y)$

$$var\bar{X} = \frac{\sigma^2}{n}$$

- $n = 1 \rightarrow var\,\overline{X_1} = \sigma^2$
- větší $n \rightarrow$ menší rozptyl $\bar{X}$
- problém: σ^2 většinou neznáme

Odvození pro výběrový průměr:

(c) Směrodatná odchylka výběrového průměru:

$$S.E.(\bar{X}) = \sqrt{var\ \bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- říkáme jí **střední chyba průměru** [standard error of mean, SEM]
- často se uvádí ve výsledcích článků
- charakterizuje „přesnost“ odhadu (pozor: přesnost odhadu ve smyslu střední kvadratické chyby (viz dále) zahrnuje i vychýlení odhadu)
- platí: čím větší výběr (n), tím přesnější odhad
- *SEM* závisí na parametru σ, který většinou neznáme a nahrazujeme ho vhodným odhadem, např. výběrovým rozptylem (za chvilku). Slovní označení „střední chyba“ se používá i tehdy, když místo σ použiji odhad.

Bodový odhad variance – výběrový rozptyl

K odhadu variability hodnot v populaci nejčastěji používáme

výběrový rozptyl $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ [sample variance]

- střední hodnota výběrového rozptylu:
  $ES^2 = \sigma^2$
- rozptyl výběrového rozptylu běžně nepotřebujeme, proto neuvádím

- výběrový momentový rozptyl $S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ většinou nepoužíváme, protože o $\frac{1}{n}$ podhodnocuje skutečný parametr σ^2 (dále)

Vsuvka – jiný tvar výběrového rozptylu:

užitečný tvar pro „ruční“ výpočet, používá se v algoritmech (je rychlejší):

$$\boldsymbol{S^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left({X_i}^2 - 2\bar{X}X_i + \bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - 2\bar{X}\sum_{i=1}^{n}X_i + n\bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - 2\bar{X}\cdot n\frac{\sum X_i}{n} + n\bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}{X_i}^2 - 2\bar{X}\cdot n\bar{X} + n\bar{X}^2\right) = \frac{\mathbf{1}}{\boldsymbol{n-1}}\left(\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{{X_i}^2} - \boldsymbol{n\bar{X}^2}\right)$$

Odvození výpočtu střední hodnoty výběrového rozptylu

$$\boldsymbol{ES^2} = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = E\left(\frac{1}{n-1}\left(\sum_{i=1}^{n} {X_i}^2 - n\bar{X}^2\right)\right) =$$

$$E(\beta \cdot X) = \beta \cdot EX$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} E\,{X_i}^2 - n \cdot E\,\bar{X}^2\right) =$$

$$= \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n \cdot \left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \frac{1}{n-1}(n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) =$$

$$= \frac{\mathbf{1}}{\boldsymbol{n-1}} \cdot \boldsymbol{(n-1)\sigma^2} = \boldsymbol{\sigma^2}$$

→ $var\,X_i = E(X_i - EX_i)^2 = E\left({X_i}^2 - 2X_i\,EX_i + (EX_i)^2\right) = E\,{X_i}^2 - 2 \cdot EX_i \cdot EX_i + (EX_i)^2 = E\,{X_i}^2 - (EX_i)^2$

odtud: $\boldsymbol{E\,{X_i}^2 = varX_i + (EX_i)^2 = \sigma^2 + \mu^2}$

→ podobně: $var\bar{X} = E(\bar{X} - E\bar{X})^2 = \cdots = E(\bar{X})^2 - (E\bar{X})^2$

odtud: $\boldsymbol{E\bar{X}^2 = var\bar{X} + (E\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2}$

Bodový odhad populační SD – výběrová směrodatná odchylka

$$S = \sqrt{S^2}$$

- tento odhad je vychýlený, skutečnou směr. odchylku v průměru podhodnocuje, protože platí $ES < \sigma$.

Vlastnosti bodového odhadu

Nestranný odhad (nevychýlený, nezkreslený) [unbiased estimation]

- když střední hodnota odhadu = teoretickému parametru
- právě jsme měli: $\boldsymbol{E\overline{X} = \mu}$ a $\boldsymbol{ES^2 = \sigma^2}$
- nestranný odhad systematicky nenadhodnocuje ani nepodhodnocuje odhadovaný parametr

- příklad vychýleného odhadu – výběrový momentový rozptyl:

$$\boldsymbol{ES_n{}^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \cdots = \frac{n-1}{n}\sigma^2}$$

vychýlení značíme $B\left(\sigma^2, S_n{}^2\right) = \boldsymbol{ES_n{}^2 - \sigma^2} = \frac{n-1}{n}\sigma^2 - \frac{n}{n}\sigma^2 = -\frac{1}{n}\boldsymbol{\sigma^2}$

$\boldsymbol{S_n{}^2}$ podhodnocuje skutečný parametr $\boldsymbol{\sigma^2}$.

Vlastnosti bodového odhadu

Asymptoticky nestranný odhad

- když odhad je sice vychýlený, ale se zvyšujícím se rozsahem výběru n se vychýlení zmenšuje až k nule
- to je případ výběrového momentového rozptylu:

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

$$ES_n^2 = \frac{n-1}{n}\sigma^2$$

vychýlení $ES_n^2 - \sigma^2 = -\frac{1}{n}\sigma^2$

$$\lim_{n\to\infty} -\frac{1}{n}\sigma^2 \to -\frac{1}{\infty}\sigma^2 = 0$$

Vlastnosti bodového odhadu

Konzistentní odhad [consistent estimation]

- pokud se s rostoucím rozsahem výběru ***n*** odhad zpřesňuje
- $E(odhadu) = parametr$
- a zároveň $\lim_{n\to\infty}\left(var(odhadu)\right) = 0$

- platí např. pro výběrový průměr:

$$\boldsymbol{E\bar{X} = \mu}$$

$$\boldsymbol{var\bar{X} = \frac{\sigma^2}{n} \quad \xrightarrow[n\to\infty]{} \frac{\sigma^2}{\infty} = 0}$$

Vlastnosti bodového odhadu

Vydatný, eficientní, nejlepší nestranný odhad [efficient estimation]

- má nejmenší rozptyl mezi všemi nestrannými odhady téhož parametru

Přesnost, kvalita odhadu [quality of the estimation]

- měříme pomocí střední kvadratické chyby odhadu
- výběrová chyba odhadu: $odhad - parametr$
- zkratka MSE(odhadu) [mean squared error]
- Kromě variability zahrnuje i vychýlení odhadu. Pro nestranné odhady (vychýlení = 0) je to totéž jako var(odhadu) a potažmo S.E.(odhadu)
- $MSE(odhadu) = E(odhad - parametr)^2 = var(odhadu) + B^2(odhadu) = \\ = E(odhad - E(odhadu))^2 + (E(odhadu) - parametr)^2$
- příklad: $MSE\left(S_n{}^2\right) = E\left(S_n{}^2 - \sigma^2\right)^2 = \cdots$

Ze statistického slovníku:

Robustní = odolný

přibližně řečeno je to schopnost spočítat „spolehlivý“ výsledek, přestože jsou narušeny předpoklady testu, odhadu apod.

Konečnostní násobitel

Většinou zahrnuje náš výběr méně než 5 % jedinců z celé populace, proto můžeme takovou populaci považovat za nekonečnou.

Pokud ovšem vybíráme z menší konečné populace a rozsah výběru je větší než 5 % všech jedinců, potom výběrový průměr $\bar{X}$ zůstává nestranným odhadem populačního průměru, ale rozptyl $\bar{X}$ se poněkud zmenší. Aby byly odhadované vlastnosti $\bar{X}$ správné, je třeba rozptyl vynásobit konečnostním násobitelem $\frac{N-n}{N-1}$.

Tedy:

$$E\bar{X} = \mu \qquad \text{... to je stejné}$$

$$var\,\bar{X} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n}$$

(Citace: Zvára, Karel: Biostatistika. Karolinum, Praha 2008.)