PRAVDĚPODOBNOST A STATISTIKA, PŘEDNÁŠKA 6. 4. 2020

Doporučená literatura: Likeš J., Machek J.: Matematická statistika. SNTL, Praha 1982.
Anděl J.: Statistické metody. MatfyzPress 2019

Základy statistiky

Historická poznámka

status = participium perfectum od sistere (stavět, ale také pravidelně se opakující)

Hlavní historické zdroje

- úřední (státní, církevní,správní) evidence pro daňové a voj. účely (Egypt 2 tis. př. n. l./Čína 1 tis. př. n. l.);
- univerzitní státověda (16. st. Itálie, Německo, Francie) - hosp., polit. a geografický popis států;
- demografie/politická aritmetika - J. Graunt, W. Petty - 17. st. - podrobné demografické studie Londýna
- teorie pravděpodobnosti - zákony velkých čísel, základy mat. statistiky (od 19. st. aplikace v technických a přírodních vědách)

Dnešní pojetí statistiky - věda zjišťování, zpracování a rozboru číselných údajů shromažďovaných buď k popisu rozsáhlých souborů (deskriptivní statistika), nebo k redukci rušivých vlivů způsobených náhodnými činiteli (matematická statistika).

Dva hlavní směry:

Popisná/deskriptivní statistika - podává ve zhuštěné číselné podobě popis „pravidelností“ kvantitativní stránky hromadných jevů.

Matematická statistika - vypracovává metody založené na předpokladu, že zjišťované údaje jsou realizací náhodných veličin a účelem jejich shromažďování je bližší určení jejich rozdělení (typ, parametry). Mat. stat. je založena na teorii pravděpodobnosti a používá její pojmy.

Základní statistické pojmy

Náhodný výběr

Nechť $X_1, \dots, X_n$ jsou nezávislé náhodné veličiny mající stejné rozdělení $R(\boldsymbol{\theta})$. Potom uspořádanou n-tici $X_1, \dots, X_n$ nazýváme náhodným výběrem (pocházejícím z rozdělení $R(\boldsymbol{\theta})$; zápis $X_1, \dots, X_n \sim R(\boldsymbol{\theta})$). Číslo n nazýváme rozsahem náhodného výběru.

Poznámky

- Při výpočtech bude náhodný výběr tvořit n-tice nezávislých realizací náhodných veličin $X_1, \dots, X_n$, tj. n-tice „naměřených“ hodnot.
- Zkratku n. v. používáme ve třech podobných významech - náhodná veličina, náhodný vektor a náhodný výběr. Z kontextu však bude význam vždy jasný.

Uspořádaný náhodný výběr

Nechť $X_1, \dots, X_n$ je náhodný výběr a $X_{(1)}, \dots, X_{(n)}$ je utvořeno z $X_1, \dots, X_n$ tak, že $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. Potom $X_{(1)}, \dots, X_{(n)}$ nazýváme uspořádaným náhodným výběrem a $X_{(i)}$ $(i = 1, \dots, n)$ nazýváme i-tou pořádkovou statistikou.

Poznámky

Statistický soubor - množina prvků, které jsou předmětem statistického zkoumání a které mají určitou vlastnost, kterou sledujeme (homogenita). Základní soubor (potenciálně všechny prvky; deskriptivní statistika) × výběrový soubor (matematická statistika).

Statistická jednotka - prvky statistického souboru.

Statistický znak - sledovaná vlastnost statistických jednotek (kvantitativní × kvalitativní; nominální × ordinální, diskrétní × spojitý, ...).

Jednorozměrný × vícerozměrný stat. soubor

3 fáze stat. zkoumání:

- statistické šetření - získávání dat formou měření, pozorování apod.; má svá „oborová“ specifika;
- statistické zpracování - zpracování dat do forem číselných charakteristik, tabulek, grafů;
- statistická analýza - vyhodnocení výsledků získaných statistickým zpracováním.

3 elementární formy stat. zpracování

- statistické charakteristiky;
- statistické tabulky;
- statistické grafy.

Základní výběrové charakteristiky

Výběrové charakteristiky jsou „empirickou“ analogií číselných charakteristik náhodných veličin (střední hodnota, kvantily, rozptyl, kovariance apod.) a slouží k získání (odhadu) hodnot jim odpovídajících číselných charakteristik rozdělení, z něhož náhodný výběr pochází.

Je-li $X_1, \dots, X_n$ je náhodný výběr, resp. $(X_1, Y_1), \dots, (X_n, Y_n)$ je dvourozměrný náhodný výběr, potom mezi nejdůležitější výběrové chrakteristiky patří:

- Výběrový (aritmetický) průměr $\quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

  Poznámky
  - Výběrový průměr je empirický „protějšek“ střední hodnoty.
  - Kromě aritmetického průměru se lze běžně setkat ještě s geometrickým a harmonickým průměrem, které mají specifické oblasti využití.

- Výběrový rozptyl $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

  Poznámky
  - Výběrový rozptyl je empirický „protějšek“ rozptylu.
  - Snadnou úpravou výše uvedeného definičního vztahu dostáváme
    $$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)},$$
    který obvykle používáme k výpočtu.
  - Kromě výběrového rozptylu S^2 se lze setkat také s rozptylem, který definujeme jako průměrnou kvadratickou odchylku „naměřených“ hodnot od jejich průměru, tj.
    $$D^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \text{ resp. } D^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n^2}$$

Mezi oběma charakteristikami je kromě číselného rozdílu (není velký, s rostoucím rozsahem n.v. klesá) také rozdíl významový, který bude vysvětlen v části věnované problematice bodových odhadů.

- Výběrový variační koeficient $C = {S}/{\bar{X}}$
- Výběrový modus x^* ... nejčastější se vyskytující hodnota v náhodném výběru
- Výběrový kvantil x_α, $\alpha \in (0,1)$

  výběrový medián $x_{0,5} = \begin{cases} X_{\left({(n+1)}/{2}\right)} & , \text{ pro } n \text{ liché} \\ \frac{1}{2}\left(X_{({n}/{2})} + X_{({n}/{2}+1)}\right) & , \text{pro } n \text{ sudé} \end{cases}$
- Výběrový r-tý obecný moment $M_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$
- Výběrový r-tý centrální moment $M_r' = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$
- Výběrová kovariance $S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \frac{n\sum_{i=1}^{n} X_iY_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n(n-1)}$
- Výběrový korelační koeficient $R_{XY} = {S_{XY}}\Big/{\sqrt{S_X^2 \cdot S_Y^2}}$

Poznámky - statistické tabulky

- Tabulka rozdělení četností

  Použití pro diskrétní náhodné výběry větších rozsahů, ve kterých se hodnoty často opakují.

x_a	n_a	p_a	k_a	$x_a n_a$	$x_a^2 n_a$
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	1	⋮	⋮
- - - -	$n = \sum n_a$	$(\sum p_a = 1)$	- - - -	$\sum x_a n_a$	$\sum x_a^2 n_a$

  x_a ... vzestupně seřazené různé hodnoty vyskytující se v n.v. n_a ... (absolutní) četnost hodnoty x_a

  $p_a = {n_a}/{n}$ ... relativní četnost hodnoty x_a $k_a = \sum_{i \le a} p_i$ ... kumulativ. rel. čet. x_a

- Tabulka třídního rozdělení četností

  Použití - n.v. velkého rozsahu velký počet různých hodnot

  Napozorované hodnoty rozdělíme do r třídních intervalů (obvykle 5 - 20; např. $r = 1 + \lfloor 3{,}3 \cdot \log_{10} n \rfloor$) a takto roztříděné hodnoty uspořádáme do četnostní tabulky;

  $(t_0, t_1\rangle, (t_1, t_2\rangle, \ldots, (t_{r-2}, t_{r-1}\rangle, (t_{r-1}, t_r)$ ... zvolené třídní intervaly (možno $t_0 = -\infty$, resp. $t_r = +\infty$);

a	t_a	$\bar{t}_a$	n_a	p_a	k_a	$\bar{t}_a n_a$	$\bar{t}_a^2 n_a$
1	t_1	$\bar{t}_1$	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
r	t_r	$\bar{t}_r$	⋮	⋮	⋮	⋮	⋮
- - - -	- - - -	- - - -	$n = \sum n_a$	$(\sum p_a = 1)$	- - - -	$\sum \bar{t}_a n_a$	$\sum \bar{t}_a^2 n_a$

a ... pořadové číslo třídního intervalu — t_a ... horní hranice třídního intervalu

$\bar{t}_a = {(t_a + t_{a-1})}/{2}$ ... střed třídního intervalu — n_a ... absolutní četnost třídního intervalu

p_a, k_a ... relativní, resp. kumulativní četnosti třídního intervalu

- Korelační tabulky ... dvourozměrný četnostní tabulka

Poznámky - statistické grafy

- Histogram četností
  analogie hustoty = graf relativních (resp. absolutních) četností
  osa x ... sloupec x_a z tabulky rozdělení četností, resp. sloupec $\bar{t}_a$ z tabulky třídního rozdělení četností;
  osa y ... sloupec p_a z tabulky rozdělení četností, resp. z tabulky třídního rozdělení četností;

- Graf empirické distribuční funkce
  = graf kumulativních relativních četností
  osa x ... sloupec x_a z tabulky rozdělení četností, resp. sloupec $\bar{t}_a$ z tabulky třídního rozdělení četností;
  osa y ... sloupec k_a z tabulky rozdělení četností, resp. z tabulky třídního rozdělení četností;

  Empirická distribuční funkce náhodného výběru $X_1, \dots, X_n$

$$F_n(x) = {n_x}/{n},$$

  kde n je rozsah n.v., n_x je počet hodnot n.v. $\leq x$

Rozdělení statistik $\bar{X}, S^2$

Poznámky

- Pojem statistika se běžně používá ve dvou významech - statistika jako věda (specifikovaná v úvodu), ale také jako funkce $T = T(X_1, \dots, X_n)$ náhodného výběru $X_1, \dots, X_n$, jejíž hodnotu lze určit bez znalosti rozdělení z něhož náhodný výběr pochází. V následující části budeme slovo statistika používat prakticky výhradně v druhém smyslu, tj. jako funkce náhodného výběru.
- Je vhodné si uvědomit, že statistika $T = T(X_1, \dots, X_n)$ je funkce náhodného výběru (tj. náhodných veličin) a je tedy náhodnou veličinou, která má své rozdělení.
- Nejdůležitějšími příklady statistik jsou: $\sum_{i=1}^n X_i$, $\sum_{i=1}^n X_i^2$, $\bar{X}$, S^2

A) Náhodné výběry z $N(\mu, \sigma^2)$

Nechť $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, potom platí:

i) $\bar{X} \sim N\left(\mu, {\sigma^2}/{n}\right)$, tj. $E(\bar{X}) = \mu$, $var(\bar{X}) = {\sigma^2}/{n}$ a proto $\frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1)$

(důsledek skutečnosti, že konvoluce normálních rozdělení má normální rozdělení, navíc

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

$var(\bar{X}) = var\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n var(X_i) = {\sigma^2}/{n}$ (využíváme nezávislost)

ii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$

iii) $\frac{\bar{X}-\mu}{S}\sqrt{n} \sim t(n-1)$

iv) Nechť $X_1, \dots, X_m \sim N(\mu_X, \sigma_X^2)$, $Y_1, \dots, Y_n \sim N(\mu_Y, \sigma_Y^2)$ jsou dva nezávislé náhodné výběry, potom

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim F(m-1, n-1).$$

Mají-li navíc oba výběry stejné rozptyly (tj. $\sigma_X^2 = \sigma_Y^2$), platí

$$\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{(m-1)S_X^2+(n-1)S_n^2}} \sqrt{\frac{mn(m+n-2)}{m+n}} \sim t(m+n-2).$$

B) Náhodné výběry z rozdělení se střední hodnotou μ a konečným rozptylem σ^2
Na náhodné výběry větších rozsahů (např. $n \geq 10$ u symetrických rozdělení, resp. $n \geq 20$ jinde) lze aplikovat centrální limitní věty, tedy $\bar{X} \sim N\left(\mu, \sigma^2/n\right)$.