MATEMATICKÁ STATISTIKA

Dana Černá

http://www.fp.tul.cz/kmd/

Katedra matematiky a didaktiky matematiky

Technická univerzita v Liberci

Matematická statistika

Matematická statistika se zabývá matematickým zpracováním dat a rozborem získaných výsledků.

Data chápeme jako realizace náhodných veličin. Statistický soubor je množina všech sledovaných objektů, jednotlivé objekty nazýváme statistické jednotky. Počet všech prvků souboru se nazývá rozsah souboru.
Rozlišujeme statistický soubor

- základní - množina všech objektů se sledovanou vlastností
- výběrový - je podmnožinou základního souboru, obsahuje objekty, ke kterým známe hodnoty sledovaných vlastností

Sledovanou vlastnost nazýváme statistický znak. Rozlišujeme statistický znak

- kvantitativní - je vyjádřen číslem
- kvalitativní - není vyjádřen číslem

(Absolutní) četnost znaku je definována jako počet statistických jednotek se stejnou hodnotou znaku. Označme absolutní četnost hodnoty ξ_i symbolem n_i, platí $\sum_{i=1}^{m} n_i = n$, kde n je rozsah souboru a m je počet všech hodnot.

Relativní četnost hodnoty ξ_i je definována vztahem $v_i = \frac{n_i}{n}$. Platí $\sum_{i=1}^{m} v_i = 1$.

Pokud statistický znak nabývá příliš mnoha hodnot, rozdělíme hodnoty do intervalů a potom používáme intervalové rozdělení četnosti. Intervaly mají zpravidla stejnou délku. Jejich počet určíme pomocí Sturgesova vzorce

$$k = 1 + 3,3 log_{10} n,$$

kde k značí počet intervalů a n označuje rozsah souboru. Daný interval potom zpravidla reprezentuje jeho střed.

Charakteristiky statistického souboru

Budeme rozlišovat

- charakteristiky polohy,
- charakteristiky variability,
- charakteristiky statistické závislosti.

Značení:

Data budeme označovat $x_1, x_2, \ldots, x_n$.

Hodnoty znaku x budeme značit $\xi_1, \xi_2, \ldots, \xi_m$ a jejich četnosti označíme $n_1, n_2, \ldots, n_m$.

Vzestupně uspořádaná data označíme $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.

Charakteristiky polohy

1. Aritmetický průměr

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ nebo } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} n_i \xi_i.$$

2. Harmonický průměr

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} \text{ nebo } \bar{x}_H = \frac{n}{\sum_{i=1}^{n}\frac{n_i}{\xi_i}}.$$

Například průměrná rychlost je harmonickým průměrem rychlostí na jednotlivých úsecích.

3. Geometrický průměr

$$\bar{x}_G = \sqrt[n]{x_1 x_2 \dots x_n} = \sqrt[n]{\xi_1^{n_1}\xi_2^{n_2}\dots\xi_m^{n_m}}.$$

Používá se k výpočtu průměrné inflace, průměrného úroku, průměrného růstu výroby, HDP apod.

4. Medián je prostřední hodnota mezi uspořádanými daty.

$Med = x_{\left(\frac{n+1}{2}\right)}$, pokud n je liché.

$$Med = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, \text{ pokud } n \text{ je sudé.}$$

5. Modus je hodnota s největší četností.

6. Dolní kvartil

$$x_{0,25} = \frac{x_{([\frac{n}{4}])} + x_{([\frac{n}{4}]+1)}}{2}, \text{ kde } [\,] \text{ označuje celou část.}$$

7. Horní kvartil

$$x_{0,75} = \frac{x_{([\frac{3n}{4}])} + x_{([\frac{3n}{4}]+1)}}{2}$$

Charakteristiky variability

1. Rozsah souboru

$$R = x_{(n)} - x_{(1)}.$$

2. Mezikvartilové rozpětí

$$Q = x_{0,75} - x_{0,25}.$$

3. Výběrový rozptyl

$$\begin{aligned} s_1^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \frac{n}{n-1}\bar{x}^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{m} n_i\,(\xi_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{m} n_i\xi_i^2 - \frac{n}{n-1}\bar{x}^2 \end{aligned}$$

$$
\begin{aligned}
s_2^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 \\
&= \frac{1}{n}\sum_{i=1}^{m}n_i\,(\xi_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{m}n_i\xi_i^2 - \bar{x}^2
\end{aligned}
$$

4. Výběrová směrodatná odchylka

$$
s_1 = \sqrt{s_1^2},\ \text{nebo}\ s_2 = \sqrt{s_2^2}
$$

5. Variační koeficient

$$
v = \frac{s_1}{\bar{x}},\ \text{nebo}\ v = \frac{s_2}{\bar{x}}
$$

Charakteristky statistické závislosti

U každého objektu budeme sledovat znak x a znak y. Potom jsou data dána ve dvojicích, dvojice (x_i, y_i) určuje zjištěné hodnoty pro i-tý objekt, nebo jsou data zadána pomocí hodnot (ξ_i, η_i) a jejich četností n_i.

Výběrová kovariance

$$\begin{aligned} cov\,(x, y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \\ &= \frac{1}{n}\sum_{i=1}^{m} n_i(\xi_i - \bar{x})(\eta_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{m} n_i \xi_i \eta_i - \bar{x}\bar{y} \end{aligned}$$

Výběrový korelační koeficient

$$\rho\left(x, y\right) = \frac{cov\left(x, y\right)}{s_x s_y},$$

kde

$$\begin{aligned} s_x &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} n_i\left(\xi_i - \bar{x}\right)^2}, \\ s_y &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{m} n_i\left(\eta_i - \bar{y}\right)^2} \end{aligned}$$

O znacích x a y řekneme, že jsou nezávislé, jestliže náhodné veličiny reprezentující tyto znaky jsou nezávislé.

Jestliže jsou znaky nezávislé, potom $\rho(x, y) \approx 0$.

Jestliže $\rho(x, y)$ není přibližně nulový, potom znaky x a y nejsou nezávislé.

Jestliže $|\rho(x, y)| \approx 1$, potom y závisí lineárně na x, tj. existují konstanty a a b takové, že $y = ax + b$.

Bodový odhad

Jestliže $X_1, X_2, \dots X_n$ jsou nezávislé stejně rozdělené náhodné veličiny, řekneme, že $X_1, X_2, \dots X_n$ je náhodný výběr. Realizace těchto náhodných veličin $x_1, x_2, \dots x_n$ nazýváme realizovaný náhodný výběr.

Předpokládejme, že $X_1, X_2, \dots X_n$ je náhodný výběr z rozdělení, které závisí na parametru $\theta \in M$.

Libovolné $\hat{\theta} \in M$ se nazývá bodový odhad parametru θ.

$\hat{\theta} \in M$ se nestranný odhad parametru θ, jestliže $E\hat{\theta} = \theta$.

VĚTA: Jestliže $X_1, X_2, \ldots X_n$ je náhodný výběr a $EX_i = \mu \in \mathbb{R}$, potom $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je nestranný odhad parametru μ.

Důkaz: $E\bar{X} = E\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$.

VĚTA: Jestliže $X_1, X_2, \ldots X_n$ je náhodný výběr a $var\ X_i = \sigma^2 \in \mathbb{R}$, potom $s^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$ je nestranný odhad parametru σ^2.

Maximálně věrohodný odhad

Předpokládejme, že $\mathbf{X} = (X_1, X_2, \ldots X_n)$ je náhodný výběr z rozdělení, které závisí na parametru $\theta \in M$, $\mathbf{x} = (x_1, x_2, \ldots x_n)$ je realizace tohoto náhodného výběru.

Definujme věrohodnostní funkci předpisem

$$L_\theta\left(\mathbf{x}\right) = P_\theta\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right),$$

pokud X_i mají diskrétní rozdělení pravděpodobnosti P_θ, a předpisem

$$L_\theta\left(\mathbf{x}\right) = \Pi_{i=1}^{n} f_\theta\left(x_i\right),$$

kde f_θ je hustota X_i, pokud X_i mají spojité rozdělení pravděpodobnosti s parametrem θ.

Maximálně věrohodný odhad parametru θ je definován:

$$\hat{\theta} = \operatorname{argmax}_{\theta \in M}\ L_\theta\left(\mathbf{x}\right) = \operatorname{argmax}_{\theta \in M}\ \ln L_\theta\left(\mathbf{x}\right).$$