PRAVDĚPODOBNOST A STATISTIKA, PŘEDNÁŠKA 4. 5. 2020

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

Kontingenční tabulka - test nezávislosti „chí-kvadrát“

Kontingenční tabulka typu (r, s), kde $r, s \geq 2$ je v podstatě dvourozměrná četnostní tabulka a používáme ji v situaci, kdy na statistických jednotkách sledujeme dva kategoriální znaky A, B, přičemž znak A může nabývat hodnot $A_1, \dots, A_r$ a znak B může nabývat hodnot $B_1, \dots, B_s$. Kontingenční tabulku obvykle zapisujeme ve tvaru uvedeném níže.

	$\boldsymbol{B_1}$	$\boldsymbol{B_2}$	$\cdots$	$\boldsymbol{B_s}$	$\sum$
$\boldsymbol{A_1}$	n_{11}	n_{12}	$\cdots$	n_{1s}	$\boldsymbol{n_{1\cdot}}$
$\boldsymbol{A_2}$	n_{21}	n_{22}	$\cdots$	n_{2s}	$\boldsymbol{n_{2\cdot}}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
$\boldsymbol{A_r}$	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$\boldsymbol{n_{r\cdot}}$
$\sum$	$\boldsymbol{n_{\cdot 1}}$	$\boldsymbol{n_{\cdot 2}}$	$\cdots$	$\boldsymbol{n_{\cdot s}}$	$\boldsymbol{n}$

kde n_{ij} je četnost (počet) jednotek, které mají současně vlastnosti A_i a B_j,
$n_{i\cdot}$ je četnost jednotek, které mají vlastnost A_i, tj. $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$,
$n_{\cdot j}$ je četnost jednotek, které mají vlastnost B_j, tj. $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$,
n je počet všech jednotek, tj. $n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{s} n_{\cdot j}$.
(četnosti $n_{i\cdot}, n_{\cdot j}$ nazýváme marginální četnosti)

Test nezávislosti v kontingenční tabulce

H_0: znaky A, B jsou nezávislé × H_1: znaky A, B nejsou nezávislé

Postup:

1. Sestavíme kontingenční tabulku (včetně marginálních četností).
2. Vypočteme testovou statistiku $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2 \Big/ \left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)$
3. Hypotézu H_0 o nezávislosti znaků A, B zamítneme na hladině α, jestliže $\chi^2 \geq \chi^2_{1-\alpha}\big((r-1)(s-1)\big)$.

Poznámky

- Doporučuje se, aby pro četnosti platilo $\frac{n_{i\cdot} n_{\cdot j}}{n} \geq 5$ $(i = 1, \dots, r; j = 1, \dots, s)$.
- Výpočet provádíme obvykle tak, že postupně vyčíslujeme sčítance $V_{ij} = \left(n_{ij} - \frac{n_{i\cdot} n_{\cdot j}}{n}\right)^2 \Big/ \left(\frac{n_{i\cdot} n_{\cdot j}}{n}\right)$, které doplňujeme do tabulky. Pokud součet již vyčíslených sčítanců V_{ij} (≥ 0) je větší nebo roven kvantilu $\chi^2_{1-\alpha}\big((r-1)(s-1)\big)$, tak výpočet ukončíme a nezávislost zamítáme.

- Statistika χ^2 je pouze testovou statistikou a není mírou závislosti, jako např. korelační koeficient. (úvaha - je zřejmé, že závislost není „důsledkem“ rozsahu náhodného výběru; pokud ale k-krát zvětšíme rozsah náhodného výběru, potom se i četnosti zvětší přibližně k-krát a tedy hodnota testové statistiky χ^2 se zvětší také zhruba k-krát → nemůže jít o míru závislosti)
- Nejjednodušší variantou kontingenční tabulky je čtypolní tabulka, kdy každý znak má pouze dvě úrovně, tj. $r = 2, s = 2$. Čtyřpolní tabulka tak má tvar

	$\boldsymbol{A_1}$	$\boldsymbol{A_2}$	$\sum$
$\boldsymbol{B_1}$	n_{11}	n_{12}	$\boldsymbol{n_{1\cdot}}$
$\boldsymbol{B_2}$	n_{21}	n_{22}	$\boldsymbol{n_{2\cdot}}$
$\sum$	$\boldsymbol{n_{\cdot 1}}$	$\boldsymbol{n_{\cdot 2}}$	$\boldsymbol{n}$

  Testová statistika má tvar $\chi^2 = \frac{n(n_{11}n_{22}-n_{21}n_{12})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$, přičemž hypotézu o nezávislosti zamítame na hladině α, jestliže $\chi^2 \geq \chi^2_{1-\alpha}(1)$.
- „Zdůvodnění“ testu nezávislosti v kontingenční tabulce
  Označíme-li X_{ij} $(i = 1, \dots, r; j = 1, \dots, s)$ náhodné veličiny vyjadřující počet jednotek (z celkového počtu n), které mají vlastnost (A_i, B_j), potom zřejmě $(X_{11}, \dots, X_{rs})$ má rs rozměrné multinomické rozdělení s parametry $n, p_{11}, \dots, p_{rs}$ a tedy veličina $\sum_{i=1}^{r}\sum_{j=1}^{s} \left(X_{ij} - np_{ij}\right)^2 / \left(np_{ij}\right)$ má rozdělení $\chi^2(rs-1)$. Označíme-li marginální pravděpodobnosti $p_{i\cdot} = \sum_{j=1}^{s} p_{ij}$ a $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$, lze hypotézu o nezávislosti psát ve tvaru H_0: $\forall i, j\ p_{ij} = p_{i\cdot} \cdot p_{\cdot j}$. V praxi jsou pravděpodobnosti p_{ij} (a tedy i $p_{i\cdot}, p_{\cdot j}$) neznámé, proto je nahradíme odhady $\hat{p}_{i\cdot} = \sum_{j=1}^{s} X_{ij} / n$, $\hat{p}_{\cdot j} = \sum_{i=1}^{r} X_{ij} / n$. Vzhledem k podmínkám $\sum_{i=1}^{r} \hat{p}_{i\cdot} = 1$, $\sum_{j=1}^{s} \hat{p}_{\cdot j} = 1$ odhadujeme pouze $(r-1) + (s-1) = r + s - 2$ parametrů. Proto zřejmě dostáváme

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \left(X_{ij} - n\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right)^2 / \left(n\hat{p}_{i\cdot}\hat{p}_{\cdot j}\right) \sim \chi^2\left(rs - 1 - (r + s - 2)\right) = \chi^2\left((r-1)(s-1)\right).$$

Regresní analýza

Funkční závislost: $y = f(x_1, \dots, x_r)$, kde $x_1, \dots, x_r$ jsou nezávisle proměnné, y závisle proměnná
(ke každému $(x_1, \dots, x_r)$ existuje nejvýše jedno $y = f(x_1, \dots, x_r)$)

V praxi však vztah mezi veličinami nemusí být (z celé řady důvodů, např. chyby měření, „opominutí“ nezávisle proměnných atd.) přímo funkční, tj. hodnotám $x_1, \dots, x_r$ mohou odpovídat různé hodnoty závisle proměnné. V tomto případě může být vhodným nástrojem pro popis závislosti regresní analýza, která vychází z toho, že závisle proměnná Y je náhodná veličina (proto velké Y), kterou lze popsat vztahem

$$Y = f(x_1, \dots, x_r) + e,$$

kde Y je náhodná veličina, tzv.vysvětlovaná proměnná,
$x_1, \dots, x_r$ jsou tzv. regresory (vysvětlující proměnné),
$f(x_1, \dots, x_r)$ je tzv. regresní funkce,
e je náhodná odchylka postihující vliv činitelů nezahrnutých mezi regresory (chyby měření apod.), navíc v našem případě $e \sim N(0, \sigma^2)$.

Cíl - na základě naměřených/napozorovaných dat (tj. hodnot vysvětlujících proměnných a jim odpovídajících hodnot vysvětlované proměnné) odhadnout regresní funkci f.

(umožňuje např. „předpovídat“ hodnoty vysvětlující proměnné Y pro „libovolné“ hodnoty regresorů $x_1, \dots, x_r$)

Jak - nejčastěji se k odhadu (parametrů) regresní funkce používá metoda nejmenších čtverců, kdy jako odhad parametrů regresní funkce zvolíme takové hodnoty, které minimalizují součet kvadrátů odchylek mezi naměřenou hodnotou vysvětlující proměnné Y_i $(i = 1, \dots, n)$ a teoretickou hodnotou regresní funkce $f(x_{i1}, \dots, x_{ir})$ pro dané hodnoty regresorů $x_{i1}, \dots, x_{ir}$ $(i = 1, \dots, n)$, tj. minimalizují

$$\sum_{i=1}^{n} \left(Y_i - f(x_{i1}, \dots, x_{ir})\right)^2$$

Jednoduchá lineární regrese (s jednou vysvětlující proměnnou)

Cíl - naměřenými daty $(x_1, Y_1), \dots, (x_n, Y_n), n \geq 3$ proložte přímku.
Vzhledem k tomu, že je velmi pravděpodobné, že všechna naměřená data neleží na jedné přímce, vzniká otázka, která přímka bude „nejlépe“ prokládat naměřená data. V souladu s výše uvedeným použijeme takovou přímku $\beta_0 + \beta_1 x$, pro kterou platí, že součet čtverců odchylek naměřených hodnot Y_i $(i = 1, \dots, n)$ a hodnot ležících na přímce příslušných x_i $(i = 1, \dots, n)$, tj. $\beta_0 + \beta_1 x_i$ je nejmenší.

Regresní model: $Y_i = \beta_0 + \beta_1 x_i + e_i$,
kde x_i $(i = 1, \dots, n)$ je hodnota vysvětlující proměnné,
Y_i je hodnota vysvětlované proměnné odpovídající hodnotě vysvětlující proměnné x_i,
β_0, β_1 jsou neznámé regresní parametry přímky (chceme odhadnout),
e_i $(i = 1, \dots, n)$ náhodná odchylka splňující podmínky $e_i \sim N(0, \sigma^2)$, kde σ^2 je neznámé (chceme odhadnout) a navíc $\forall i \neq j \; cov(e_i, e_j) = 0$.

Neznáme regresní parametry β_0, β_1 odhadneme (odhady označíme b_0, b_1) metodou nejmenších čtverců následovně - minimalizujeme funkci dvou proměnných

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(Y_i - (\beta_0 + \beta_1 x_i)\right)^2,$$

tedy řešíme soustavu

$$\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} = 0, \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} = 0,$$

resp. po rozepsání dostáváme soustavu normálních rovnic

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i, \; \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i Y_i.$$

Řešením jsou následující lineární NNO regresních parametrů:

$$b_1 = \frac{n \sum_{i=1}^{n} x_i Y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \; b_0 = \frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{Y} - b_1 \bar{x}.$$

$$\left(\bar{Y} = \sum_{i=1}^{n} Y_i / n \text{ a } \bar{x} = \sum_{i=1}^{n} x_i / n\right)$$

$\hat{Y} = b_0 + b_1 x$ je tzv. přímka odhadu ($\hat{Y}$ je nestranný odhad hodnoty vysvětlované proměnné odpovídající hodnotě vysvětlující proměnné x).

Poznámky

- $\hat{e_i} = Y_i - \hat{Y}_i$ ... tzv. reziduum
  $S_t = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ ... tzv. celkový součet čtverců
  $S_r = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$ ... tzv. regresní součet čtverců
  $S_e = \sum_{i=1}^{n} \hat{e_i}^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ ... tzv. reziduální součet čtverců

- Platí: i) $S_t = S_r + S_e$

  ii) $s_e^2 = \frac{1}{n-2} S_e = \left(\sum_{i=1}^n Y_i^2 - b_0 \sum_{i=1}^n Y_i - b_1 \sum_{i=1}^n x_i Y_i\right) \Big/ (n-2)$

  je nestranný odhad parametru σ^2, tzv. reziduální rozptyl

- $R^2 = {}^{S_r}/_{S_t} = 1 - {}^{S_e}/_{S_t}$ je tzv. koeficient determinace (zřejmě $0 \leq R^2 \leq 1$), který vyjadřuje nakolik je daný model schopen vysvětlit výběrový rozptyl vysvětlované proměnné Y ($R^2 \geq 0{,}95$ … časté kritérium pro přijetí modelu).

Kvadratická regrese - proložení paraboly

Cíl - naměřenými daty $(x_1, Y_1), \dots, (x_n, Y_n), n \geq 4$ proložte parabolu.

Regresní model: $Y_i = \beta_0 + \beta_1 x_i + \beta_2 {x_i}^2 + e_i$,

$\beta_0, \beta_1, \beta_2$ jsou neznámé regresní parametry (chceme odhadnout),

(přesto, že prokládáme parabolu, jde o lineární regresi, neboť závislost na parametrech je lineární)

Analogickým postupem, tj.minimalizací funkce $Q(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^n \left(Y_i - (\beta_0 + \beta_1 x_i + \beta_2 {x_i}^2)\right)^2$ dostáváme následující odhady regresních parametrů

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y},$$

kde $\boldsymbol{X} = \begin{pmatrix} 1 & x_1 & {x_1}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & {x_n}^2 \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, tj. $\boldsymbol{X}^T \boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 \\ \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i^3 & \sum_{i=1}^n x_i^4 \end{pmatrix}, \boldsymbol{X^T Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i Y_i \\ \sum_{i=1}^n {x_i}^2 Y_i \end{pmatrix}$.

Poznámka - vícenásobná regrese

Kromě regrese s jednou vysvětlující proměnnou je samozřejmě využívána i vícenásobná regrese, kdy vysvětlovaná proměnná závisí na více vysvětlujících proměnných. Jednoduchým příkladem je úloha, kdy daty $\left(x_1^{(1)}, x_1^{(2)}, Y_1\right), \dots, \left(x_n^{(1)}, x_n^{(2)}, Y_n\right), n \geq 4$ (tj. máme dvě vysvětlující proměnné $x^{(1)}, x^{(2)}$) prokládáme rovinu.

Regresní model: $Y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + e_i$

Analogický výše uvedenému dostáváme následující odhady regresních parametrů

$$\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y},$$

kde $\boldsymbol{X} = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & x_n^{(2)} \end{pmatrix}, \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, tj.

$$\boldsymbol{X}^T \boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i^{(1)} & \sum_{i=1}^n x_i^{(2)} \\ \sum_{i=1}^n x_i^{(1)} & \sum_{i=1}^n \left(x_i^{(1)}\right)^2 & \sum_{i=1}^n x_i^{(1)} x_i^{(2)} \\ \sum_{i=1}^n x_i^{(2)} & \sum_{i=1}^n x_i^{(1)} x_i^{(2)} & \sum_{i=1}^n \left(x_i^{(2)}\right)^2 \end{pmatrix}, \boldsymbol{X^T Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_i^{(1)} Y_i \\ \sum_{i=1}^n x_i^{(2)} Y_i \end{pmatrix}.$$