Matematická statistika

Šárka Hudecová

Katedra pravděpodobnosti a matematické statistiky
Matematicko-fyzikální fakulta Univerzity Karlovy

letní semestr 2012

Opakování

Opakování: Testy o střední hodnotě normálního rozdělení

1. jednovýběrový t-test
2. párový t-test
3. dvouvýběrový t-test

Testy v binomickém rozdělení

- jednovýběrová situace
- dvouvýběrová situace

Neparametrické testy

- jednovýběrový Wilcoxonův test
- dvouvýběrový Wilcoxonův test

Test pravděpodobnosti v binomickém rozdělení

Situace: $Y \sim \mathrm{Bi}(n, p)$, chceme testovat na hladině α

$$H_0 : p = p_0 \quad \text{proti} \quad H_1 : p \neq p_0$$

(resp. proti jednostranným alternativám)

Testová statistika

$$Z = \frac{Y - np_0}{\sqrt{np_0(1 - p_0)}}$$

- z CLV plyne, že za H_0 má Z přibližně $\mathrm{N}(0, 1)$ rozdělení

Test: Hypotézu H_0 zamítáme ve prospěch

- $H_1 : p \neq p_0$, pokud $|Z| > z_{1-\alpha/2}$
- $H_1 : p > p_0$, pokud $Z > z_{1-\alpha}$
- $H_1 : p < p_0$, pokud $Z < -z_{1-\alpha}$

Poznámky

- asymptotický test na hladině α
- pomocí Z lze odvodit interval spolehlivosti pro p
- někdy korekce pro spojitost (Yatesova korekce)

$$\widetilde{Z} = \frac{|Y - np_0| - 0.5}{\sqrt{np_0(1 - p_0)}} \cdot \mathrm{sgn}(Y - np_0)$$

  pak za H_0 má $\widetilde{Z} \stackrel{\sim}{\sim} \mathrm{N}(0, 1)$
- existuje i přesný postup bez použití aproximací

Příklad — náboženské vyznání

Příklad

- sčítání lidu v roce 1991: 39.2 % obyvatel římskokatolické (ŘK) vyznání
- zajímá nás, zda je podíl vysokoškolských učitelů s tímto vyznáním stejný nebo odlišný
- průzkum: 62 VŠ učitelů, z nich 30 ŘK vyznání

Model:

- Y je počet „úspěchů" = počet učitelů s ŘK vyznáním z n náhodně vybraných
- $Y \sim \mathrm{Bi}(n, p)$

Chceme testovat

$$H_0 : p = 0.392 \quad \text{proti} \quad H_1 : p \neq 0.392$$

Příklad — řešení

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

- $Y = 30,\ n = 62,\ \alpha = 0.05$
- bodový odhad $\hat{p} = \frac{30}{62} = 0.484$
- testová statistika

$$Z = \frac{30 - 62 \cdot 0.392}{\sqrt{62 \cdot 0.392 \cdot (1 - 0.392}} = 1.482$$

test: $|Z| = 1.482 < z(0.975) = 1.960 \rightsquigarrow$ nelze zamítnout H_0 a tedy nelze prokázat, že by byl podíl ŘK mezi VŠ učiteli odlišný od celé populace

- korekce pro spojitost:

$$\tilde{Z} = \frac{|30 - 62 \cdot 0.392| - 0.5}{\sqrt{62 \cdot 0.392 \cdot (1 - 0.392}} \cdot \mathrm{sgn}(30 - 62 \cdot 0.392) = 1.352$$

opatrnější výsledek

Řešení v programu R

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

Program R: funkce `prop.test` počítá Z^2, resp. $\widetilde{Z}^2$

```
> prop.test(30,62,p=0.392,correct=F)
1-sample proportions test without continuity correction
data: 30 out of 62, null probability 0.392
X-squared = 2.1956, df = 1, p-value = 0.1384
alternative hypothesis: true p is not equal to 0.392
95 percent confidence interval:
0.3640986 0.6055254
sample estimates:
p 0.483871
```

Test homogenity dvou binomických rozdělení

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

Situace: $X \sim \mathrm{Bi}(m, p_1)$, $Y \sim \mathrm{Bi}(n, p_2)$ nezávislé, chceme testovat

$$H_0 : p_1 = p_2 \quad \text{proti} \quad H_1 : p_1 \neq p_2$$

(resp. proti jednostranným alternativám)

Příklady

- procentuální výskyt kuřáků mezi muži a ženami
- volební preference určité strany mezi lidmi s VŠ a bez VŠ
- úspěšnost experimentálního a standardního léku

Konstrukce testové statistiky

- označíme $x = \frac{X}{m}$ a $y = \frac{Y}{n} \rightsquigarrow x - y$ odhaduje $p_1 - p_2$
- vezmeme-li

$$U = \frac{x - y - (p_1 - p_2)}{\sqrt{\operatorname{var}(x-y)}} = \frac{x - y - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}},$$

  pak lze ukázat, že U má přibližně $N(0,1)$
- ve jmenovateli je třeba nahradit neznámé p_1 a p_2 jejich odhady
  za $H_0 : p_1 = p_2 \rightsquigarrow$ vezmeme společný odhad $z = \frac{X+Y}{m+n}$
- testová statistika

$$Z = \frac{x - y}{\sqrt{z(1-z)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

  má za H_0 přibližně $N(0,1)$ rozdělení

Test homogenity dvou binomických rozdělení

Testová statistika

$$Z = \frac{x - y}{\sqrt{z(1-z)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

má za H_0 přibližně $N(0,1)$ rozdělení

Test: Hypotézu $H_0 : p_1 = p_2$ zamítáme ve prospěch

- $H_1 : p_1 \neq p_2$, pokud $|Z| > z_{1-\alpha/2}$
- $H_1 : p_1 > p_2$, pokud $Z > z_{1-\alpha}$
- $H_1 : p_1 < p_2$, pokud $Z < -z_{1-\alpha}$

Poznámky

- existují i jiné postupy a modifikace testové statistiky Z
- je možné uvažovat obecnější hypotézu $H_0 : p_1 - p_2 = \delta_0$ nebo $H_0 : \frac{p_1}{p_2} = r_0$
- lze konstruovat interval spolehlivosti pro $p_1 - p_2$ nebo pro $\frac{p_1}{p_2}$
- lze pracovat s korekcí pro spojitost

Příklad – zuby

Příklad

Úkolem je zjistit, zda teplotní šoky poškozují zubní sklovinu a narušují tak mechanickou odolnost zubů.
Studie:

- 50 zubů $\rightsquigarrow$ teplotní šoky $\rightsquigarrow$ 21 zlomených
- 50 zubů $\rightsquigarrow$ kontrolní skupina $\rightsquigarrow$ 11 zlomených

Model:

- $X = \#$ zlomených zubů vystavených teplotním šokům, $Y = \#$ zlomených kontrolních zubů
- $X \sim \mathrm{Bi}(n, p_1)$, $Y \sim \mathrm{Bi}(m, p_2)$

Chceme testovat

$$H_0 : p_1 = p_2 \quad \text{proti } H_1 : p_1 > p_2$$

Příklad – řešení

- $n = m = 50,\ X = 21,\ Y = 11$
- $x = 0.42,\ y = 0.22,\ z = 0.32$
- testová statistika

$$Z = \frac{0.42 - 0.22}{\sqrt{0.32(1-0.32)\left(\frac{1}{50}+\frac{1}{50}\right)}} = 2.144$$

- $Z = 2.144 > z(0.95) = 1.645 \rightsquigarrow$ zamítáme H_0 a prokázali jsme, že teplotní šoky negativně ovlivňuji mechanickou odolnost zubů

Řešení v programy R


```
> prop.test(c(21,11),c(50,50),correct=F,
   alternative=‘‘greater’’)
2-sample test for equality of proportions without
continuity correction
data: c(21, 11) out of c(50, 50)
X-squared = 4.5956, df = 1, p-value = 0.01603
alternative hypothesis: greater
95 percent confidence interval:
0.0501106 1.0000000
sample estimates:
prop 1 prop 2
0.42 0.22
```

Neparametrické testy

t-testy:

- hypotézy o populačním průměru (střední hodnoty)
- předpoklad normality

Neparametrické testy

- není nutné specifikovat přesný typ rozdělení
- často jsou založeny na tzv. pořadích $\rightsquigarrow$ tzv. pořadové testy
- obecnější hypotézy — rovnost mediánů, shoda celých rozdělení apod.
- v případě normálního rozdělení menší síla ve srovnání s t-testy

Pořadí

- pozorování $Y_1, \ldots, Y_n$ ze spojitého rozdělení
- uspořádaný výběr (varianční řada)

$$Y_{(1)} \leq \ldots Y_{(n)}$$

  hodnoty uspořádané dle velikosti od minima po maximum
- pořadí: $R_1, \ldots, R_n$
  $R_i \leftrightsquigarrow$ kterém místě v uspořádaném výběru veličina Y_i stojí
  (v případě shod se berou průměry z odpovídajících pořadí)

Příklad

- máme hodnoty $7, 4, 5, 13, 2$
- uspořádaný výběr $2, 4, 5, 7, 13$
- pořadí $4, 2, 3, 5, 1$

Jednovýběrový Wilcoxonův test

Situace: $X_1, \dots X_n$ výběr ze spojitého rozdělení symetrického kolem bodu a (tj. $a = m_X$ medián X), chceme testovat

$$H_0 : m_X = m_0, \quad \text{proti } H_1 : m_X \neq m_0$$

Postup

- vyloučíme případy $X_i = m_0$ (a dle toho upravíme n)
- definujeme $Y_i = X_i - m_0$
- uspořádáme $|Y_i|$ dle velikosti
- R_i^+ pořadí $|Y_i|$
- za H_0 by součty R_i^+ pro kladná a záporná Y_i měly být srovnatelné
  (je-li např. součet pro kladná Y_i výrazně větší než pro záporná $\rightsquigarrow$ svědčí to pro $m_X > m_0$)

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

Jednovýběrový Wilcoxonův test

- W součet pořadí R_i^+ pro $Y_i > 0$
- **testová statistika**

$$Z = \frac{W - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

za H_0 má Z přibližně $N(0,1)$ rozdělení

Provedení testu: Hypotézu $H_0 : m_X = m_0$ zamítáme ve prospěch

- $H_1 : m_X \neq m_0$, pokud $|Z| > z_{1-\alpha/2}$
- $H_1 : m_X > m_0$, pokud $Z > z_{1-\alpha}$
- $H_1 : m_X < m_0$, pokud $Z < -z_{1-\alpha}$

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

Poznámky

- k zamítnutí H_0 může vést také výraznější porušení předpokladu symetrie (přestože třeba $m_X = m_0$)
- v případě výskytu shod $X_i = m_0 \rightsquigarrow$ lze korekce ve jmenovateli
- lze uvažovat korekci pro spojitost
- existují i přesné postupy bez využití aproximací

Matematická statistika
Testy v binomickém rozdělení
Test pravděpodobnosti v binomickém rozdělení
Test homogenity dvou binomických rozdělení
Neparametrické testy
Jednovýběrový Wilcoxonův test

Příklad — pivo

Příklad

Provádíme průzkum, jaký skutečný objem piva točí v nejmenované hospodě. Zakoupeno bylo 10 piv a jejich objem byl (v litrech):

0.510, 0.462, 0.491, 0.466, 0.461, 0.503, 0.495, 0.488, 0.512, 0.505.

Chceme otestovat, zda hostinský netočí pod míru.

Použijeme popsaný postup:

X_i	0.51	0.462	0.491	0.466	0.461	0.503	0.495	0.488	0.512	0.505
Y_i	0.01	-0.038	-0.009	-0.034	-0.039	0.003	-0.005	-0.012	0.012	0.005
$\|Y_i\|$	0.01	0.038	0.009	0.034	0.039	0.003	0.005	0.012	0.012	0.005
R_i^+	5	9	4	8	10	1	2.5	6.5	6.5	2.5

$W = 5 + 1 + 6.5 + 2.5 = 15$

Příklad — řešení

- $W = 5 + 1 + 6.5 + 2.5 = 15$
- testová statistika: $Z = -1.274$, tj. $Z > -z(0.95) = -1.645 \rightsquigarrow$ nelze zamítnou H_0

Provedení v programu R:

```
> wilcox.test(x,mu=0.5,alternative=‘‘less’’,
     exact=F,corr=F)

Wilcoxon signed rank test

data: pivo
V = 15, p-value = 0.101
alternative hypothesis: true location is less than 0.5
```