Programovanie v prostredí R

VI. testy štatistických hypotéz

Aleš Kozubík

Katedra Matematických Metód a Operačnej Analýzy

12.12.2019

Používané dáta

Budeme používať datasety `iris`, `trees`, `warpbreaks` a `PlantGrowth`, ktoré sú obsiahnuté v samotnom systéme R.

S ich obsahom sa môžeme bližšie oboznámiť pomocou funkcie `help(<meno>)`, kde `<meno>` je názov jednotlivých datasetov.

Čo si ukážeme

- Testy významnosti
- Miery asociácie dát (štatistickej závislosti)
- Štatistické testy.

Podstata testov významnosti

Overuje sa platnosť tzv. nulovej hypotézy H_0 oproti alternatívnej hypotéze H_A. Obvykle ide o tvrdenie o hodnote nejakého parametra.

Na základe náhodného výberu (experimentu) sa zostrojí tzv. testovacia štatistika.

Na základe príslušnosti vypočítanej hodnoty testovacej štatistiky do intervalu spoľahlivosti zodpovedajúceho požadovanej hladine významnosti buď zamietneme alebo nezamietneme nulovú hypotézu.

Typy testov významnosti

Existujú tri druhy testov:

1. **Jednovýberový test** na základe náhodného výberu sa overuje hypotéza o rovnosti parametra predpokladanej hodnote.
2. **Dvojvýberový test** na základe dvoch nezávislých náhodných výberov sa testuje zhoda parametrov, t.j. či pochádzajú zo súborov s rovnakými charakteristikami.
3. **Párový test** medzi jednotlivými hodnotami vo dvoch náhodných výberoch je jednoznačná korešpondecia, musia mať rovnaký rozsah.

Test významnosti pre podiel populácie

Nulová hypotéza $H_0 : p = p_0$, tj. podiel p_0 zo základného súboru má požadovanú vlastnosť.

Alternatívna hypotéza je $H_A : p \neq p_0$ pre obojstrannú alternatívu, resp. $H_A : p > p_0$ alebo $H_A : p < p_0$ pre jednostrannú alternatívu.

Ak $\hat{p}$ je odhad podielu p, tak testovacia štatistika má tvar

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

Ak je hypotéza pravdivá, má Z štandardné normálne rozdelenie $N(0, 1)$.

Test významnosti pre podiel populácie

Používame funkciu `prop.test()`. Má niekoľko bežných argumentov:

- `x` počet prvkov s danou vlastnosťou vo výbere,
- `n` rozsah náhodného výberu,
- `p` predpokladaný podiel prvkov s danou vlastnosťou,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alt` či ide o obojstranný test (alternatívna hypotéza $H_A: p \neq p_0$) alebo jednostranný test (alternatívna hypotéza $H_A: p > p_0$ hodnota `greater`, alebo $H_A: p < p_0$ hodnota `less`)

Test významnosti pre podiel populácie

Príklad

V skupine 1600 osôb bolo objavených 850 osôb s nadváhou. Overte platnosť hypotézy, že nadpolovičná časť populácie má nadváhu.

Riešenie:

```
> prop.test(x=850,n=1600,p=0.5,alt="greater")
   1-sample proportions test with continuity correction

data:  850 out of 1600, null probability 0.5
X-squared = 6.1256, df = 1, p-value = 0.006662
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5103813 1.0000000
sample estimates:
      p
0.53125
```

Test významnosti pre strednú hodnotu

Jednovýberový test hypotézy $H_0 : \mu = \mu_0$ proti alternatíve $H_A : \mu \neq \mu_0$ resp. jednostranné alternatívy $H_A : \mu > \mu_0, H_{:} \mu < \mu_0$

Testovacia štatistika má tvar

$$T = \frac{\overline{x} - \mu}{s/\sqrt{n}}.$$

Ak je hypotéza pravdivá, tak T má Studentovo rozdelenie $t(n-1)$ s $n-1$ stupňami voľnosti.

Test významnosti pre strednú hodnotu

Na vykonanie testu slúži funkcia `t.test()` s argumentmi:

- `x` premenná obsahujúca prvky náhodného výberu,
- `mu` predpokladaná stredná hodnota,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alt` či ide o obojstranný alebo jednostranný test (default je obojstranný).

Test významnosti pre strednú hodnotu

Príklad

S testovaným vozidlom bolo vykonaných 10 jázd a postupne nameraná priemerná spotreba 10.1, 12, 9.6, 10.5, 11, 9.8, 10.6, 10.7, 9.7, 11 l/100 km. Chceme overiť, či výsledky zodpovedajú deklarovanej spotrebe 10 l km

```
> x<-c(10.1,12,9.6,10.5,11,9.8,10.6,10.7,9.7,11)
> t.test(x,mu=10,conf.level=0.99,alt="greater")
        One Sample t-test
data:  x
t = 2.1429, df = 9, p-value = 0.03037
alternative hypothesis: true mean is greater than 10
99 percent confidence interval:
 9.841664      Inf
sample estimates:
mean of x
     10.5
```

Test významnosti pre rozptyl

Jednovýberový test hypotézy $H_0 : \sigma^2 = \sigma_0^2$ proti alternatíve $H_A : \sigma^2 \neq \sigma_0^2$ resp. jednostranné alternatívy $H_A : \sigma^2 > \sigma_0^2, H_A : \sigma^2 < \sigma_0^2$

Testovacia štatistika má tvar

$$\chi = \frac{(n-1) \cdot s^2}{\sigma_0^2}.$$

Ak je hypotéza pravdivá, tak χ má Pearsonovo rozdelenie $\chi^2(n-1)$ s $n-1$ stupňami voľnosti.

Test významnosti pre rozptyl

Na vykonanie testu slúži funkcia `varTest()`, ktorá je súčasťou balíčka `EnvStats`. Jej argumenty:

- `x` premenná obsahujúca prvky náhodného výberu,
- `sigma.squared` predpokladaná hodnota rozptylu,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alternative` či ide o obojstranný alebo jednostranný test (default je obojstranný).

Test významnosti pre rozptyl

Príklad

S testovaným vozidlom bolo vykonaných 10 jázd a postupne nameraná priemerná spotreba 10.1, 12, 9.6, 10.5, 11, 9.8, 10.6, 10.7, 9.7, 11 l/100 km. Chceme overiť, či je rozptyl menší ako 1.

```
> x<-c(10.1,12,9.6,10.5,11,9.8,10.6,10.7,9.7,11)
> varTest(x,sigma.squared=1,conf.level=0.99,alt="less")
Null Hypothesis:                 variance = 1
Alternative Hypothesis:          True variance is less than 1
Test Name:                       Chi-Squared Test on Variance
Estimated Parameter(s):          variance = 0.5444444
Data:                            x
Test Statistic:                  Chi-Squared = 4.9
Test Statistic Parameter:        df = 9
P-value:                         0.1570631
99% Confidence Interval:         LCL = 0.000000
                                 UCL = 2.346855
```

Znamienkový test – test významnosti pre medián

Jednovýberový test hypotézy $H_0 : \mathrm{Me}_X = m$ proti alternatíve $H_A : \mathrm{Me}_X \neq m$ resp. jednostranné alternatívy $H_A : \mathrm{Me}_X > m, H_A : \mathrm{Me}_X < m$

Testovacia štatistika má tvar

$$T = \#\{X_i, X_i > m\}.$$

Ak je hypotéza pravdivá, tak T má Binomické rozdelenie $\mathrm{Bi}(n, 1/2)$.

Znamienkový test – test významnosti pre medián

Pre výpočet testovacej štatistiky využijeme funkciu `\sum()`.

Pre určenie p-hodnôt využijeme pravdepodobnostnú funkciu binomického rozdelenia `pbinom()`.

Je však potrebné si uvedomiť, že pre $\mathbb{P}(T \geq k) = 1 - \mathbb{P}(T < k)$ musíme použiť hodnotu `1-pbinom(k-1,n,1/2)`.

Znamienkový test – príklad

Predpokladajme, že výpis telefónneho účtu obsahuje nasledujúce dĺžky hovorov v minútach:

$$2\ 1\ 3\ 3\ 3\ 3\ 1\ 3\ 16\ 2\ 2\ 12\ 20\ 31$$

Chceme overiť, či medián dĺžky hovoru je 5 minút alebo kratší.

Chceme teda testovať hypotézu

$$H_0 : \mathrm{Me} = 5 \text{ proti alternatíve } H_A : \mathrm{Me} < 5$$

Znamienkový test – príklad

```
> hovory<-c(2,1,3,3,3,3,1,3,16,2,2,12,20,3,1)
> T<-sum(hovory > 5)
> n<-length(hovory)
> k<-n-T
> 1-pbinom(k-1,n,1/2)
[1] 0.01757812
```

p-hodnota je 0.01757812, môžeme teda so spoľahlivosťou 98 % tvrdiť, že medián je nižší ako 5 minút.

Wilcoxonov test

Je neparametrickým testom pre strednú hodnotu resp. medián.

Predpokladá sa, že ide o náhodný výber zo symetrického rozdelenia, ktoré sa však výrazne odlišuje od normálneho. Vzhľadom na symetriu je Me = μ.

Jednovýberový test hypotézy $H_0 : \mathrm{Me}_X = m$ proti alternatíve $H_A : \mathrm{Me}_X \neq m$ resp. jednostranné alternatívy $H_A : \mathrm{Me}_X > m, H_A : \mathrm{Me}_X < m$ resp. to isté pre stredné hodnoty.

Testovacia štatistika má tvar

$$T = \sum_{i: X_i > m} \mathrm{poradie}(|X_i - m|).$$

Wilcoxonov test

Na vykonanie testu slúži funkcia `wilcox.test()`. Jej vybrané argumenty:

- `x` premenná obsahujúca prvky náhodného výberu,
- `mu` predpokladaná hodnota mediánu resp. strednej hodnoty,
- `conf.level` hladina spoľahlivosti (default je 0.95),
- `alt` či ide o obojstranný alebo jednostranný test (default je obojstranný),
- `conf.int` logická premenná, či bude vo výsledkoch aj interval spoľahlivosti.

Wilcoxonov test – príklad

Anděl,J.:*Statistické metody*, str. 88, príklad 7.8

10 pokusných osôb malo nezávisle od seba bez nácviku odhadnúť dobu jednej minúty od daného signálu. Boli získané tieto výsledky v sekundách:

$$53,\ 48,\ 45,\ 55,\ 63,\ 51,\ 66,\ 56,\ 50,\ 58.$$

Chceme overiť, či medián rozdelenia je 60 sekúnd.

```
> casy<-c(53,48,45,55,63,51,66,56,50,58)
> wilcox.test(casy,mu=60,alt="less")

        Wilcoxon signed rank test
data:   casy
V = 7, p-value = 0.01855
alternative hypothesis: true location is less than 60
```

Dvojvýberový test pre podiely populácie

Porovnávame zastúpenie určitých prvkov vo dvoch populáciách.

K dispozícii máme dva náhodné výbery, prvý s rozsahom n_1, druhý s rozsahom n_2.

V prvom výbere je x_1 prvkov s danou vlastnosťou, resp. podiel prvkov s danou vlastnosťou je $\hat{p}_1$, v druhom x_2 prvkov resp. podiel $\hat{p}_2$.

Testujeme hypotézu $H_0 : p_1 = p_2$ proti alternatíve $H_A : p_1 \neq p_2$ resp. jednostranné alternatívy $H_A : p_1 > p_2$ alebo $H_A : p_1 < p_2$.

Dvojvýberový test pre podiely populácie

Označme ešte ako $\hat{p}$ podiel prvkov s danou vlastnosťou v oboch výberoch dohromady, teda:

$$\hat{p} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}.$$

Testovacia štatistika má tvar

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Dvojvýberový test pre podiely populácie

V systéme R použijeme funkciu `prop.test()`, kde ako parametre použijeme `x=c(x1,x2)` a `n=c(n1,n2)`.

Na stránke
https://www.fibalivestats.com/u/SBL/986090/bs.html
je možné pozrieť štatistiky z basketbalového zápasu Svit – Žilina. chceme porovnať, či domáci mali štatisticky významne lepšiu úspešnosť trojbodovej streľby ako hostia.

Zo zápisu sme zistili, že v prípade domácich je počet streleckých pokusov $n_1 = 38$, z toho bolo $x_1 = 14$ úspešných.

Podobne v prípade hostí bolo $n_2 = 23$ pokusov, z toho bolo $x_2 = 7$ úspešných.

Dvojvýberový test pre podiely populácie

Riešenie v systéme R

```
> pokusy<-c(38,23)
> premenene<-c(14,7)
> prop.test(premenene,pokusy)
2-sample test for equality of proportions

data:  premenene out of pokusy
X-squared = 0.054028, df = 1, p-value = 0.8162
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.2134842  0.3416306
sample estimates:
   prop 1    prop 2
0.3684211 0.3043478
```

Dvojvýberový t-test pre stredné hodnoty

Máme k dispozícii dva náhodné výbery, prvý má rozsah n_1, druhý n_2, na ich základe určíme odhady strednej hodnoty ako priemery $\overline{X}$ resp. $\overline{Y}$.

Príslušné výberové rozptyly označíme s_X^2 resp. s_Y^2.

Testujeme hypotézu $H_0 : \mu_X = \mu_Y$ proti alternatíve $H_A : \mu_X \neq \mu_Y$ resp. $H_A : \mu_X > \mu_Y$ alebo $H_A : \mu_X < \mu_Y$.

Testovacia štatistika:

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1 - 1)s_X^2 + (n_2 - 1)s_Y^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \sim t(n_1 + n2 - 2).$$

Dvojvýberový *t*-test pre stredné hodnoty

Pre realizáciu testu v systéme R použijeme funkciu `t.test()`, ktorá má tieto argumenty:

- `x,y` numerické vektory, hodnoty z náhodného výberu,
- `alt` alternatívna hypotéza, môže byť `two.sided` (implicitná) alebo `less` resp. `greater` pre jednostranné alternatívy,
- `var.equal` logická premenná, udáva, či predpokladáme rovnaký rozptyl (implicitne FALSE),
- `conf.level` hladina spoľahlivosti, implicitne 0.95.

Dvojvýberový *t*-test pre stredné hodnoty

Boli namerané nasledujúce váhy pre skupinu žien (v kg):38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5 a pre skupinu mužov:67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4. Chceme porovnať, či stredná váha žien je nižšia ako u mužov.

```
> zeny<-c(38.9,61.2,73.3,21.8,63.4,64.6,48.4,48.8,48.5)
> muzi<-c(67.8,60,63.4,76,89.4,73.3,67.3,61.3,62.4)
> t.test(zeny,muzi,alt="less")
        Welch Two Sample t-test
data:  zeny and muzi
t = -2.7842, df = 13.114, p-value = 0.007692
alternative hypothesis: true difference in means is less tha
95 percent confidence interval:
      -Inf -6.153708
sample estimates:
mean of x mean of y
```

Fisherov test rovnosti rozptylov

Máme dva náhodné výbery X a Y s rozsahmi n_X a n_Y. Z nich určíme výberové rozptyly s_X^2 a s_Y^2.

Overujeme platnosť hypotézy $H_0 : \sigma_X^2 = \sigma_Y^2$ proti alternatívam $H_A : \sigma_X^2 \neq \sigma_Y^2$ resp. jednostranným $H_A : \sigma_X^2 > \sigma_Y^2$ alebo $H_A : \sigma_X^2 < \sigma_Y^2$.

Testovacia štatistika má v tomto prípade tvar

$$F = \frac{s_X^2}{s_Y^2} \sim F(n_X - 1, n_Y - 1).$$

Fisherov test rovnosti rozptylov

V systéme R je k dispozícii funkcia `var.test()` s premennými:

- `x,y` numerické vektory, hodnoty z náhodného výberu,
- `alt` alternatíva, môže byť `two.sided` resp. `less` alebo `greater`,
- `conf.level` hladina spoľahlivosti.

Fisherov test rovnosti rozptylov

V predchádzajúcich súboroch s váhami porovnáme rozptyly v oboch skupinách.

```
> var.test(zeny,muzi,alt="two.sided")
        F test to compare two variances
data:  zeny and muzi
F = 2.7675, num df = 8, denom df = 8, p-value = 0.1714
alternative hypothesis: true ratio of variances is not equal
95 percent confidence interval:
  0.6242536 12.2689506
sample estimates:
ratio of variances
          2.767478
```

Párový test

Ide o porovnávanie náhodných výberov, ktoré spolu nejakým spôsobom súvisia.

Typické napr. pre skúmanie vlastností dvojičiek, porovnávanie účinnosti nejakých procedúr, testovaním pred a po absolvovaní a pod.

K dispozícii sú teda dva náhodné výbery $X_1, X_2, \ldots, X_n$ a $Y_1, Y_2, \ldots, Y_n$. Rozdiely $X_i - Y_i$ musia byť pri zhode nezávislé a identicky rozdelené.

Párový test

Test hypotézy $H_0 : \mu_X = \mu_y$, proti alternatívam $H_A : \mu_X \neq \mu_Y$, $H_A : \mu_X > \mu_Y$ resp. $H_A : \mu_X < \mu_Y$ potom prechádza na test pre rozdiel $X - Y$:

$H_0 : \mu = 0$ proti $H_A : \mu \neq 0$ alebo $H_A : \mu > 0$ resp. $H_A : \mu < 0$.

Pri predpoklade normálneho rozdelenia použijeme t-test.

Intervaly spoľahlivosti

Smerodajná odchýlka

Príklad

Na overenie účinnosti kurzu boli overované vedomosti frekventantov pred a po absolvovaní kurzu. Dosiahnuté počty bodov sú zhrnuté do tabuľky:

Test	**Výsledné body**									
Pred	77	56	64	60	57	53	72	62	65	66
Po	88	74	83	68	58	50	67	64	74	60

Porovnajme výsledky pred a po absolvovaní kurzu, najskôr predpokladajúc, že ide o dva náhodné výbery a potom, že ide o spárované výsledky tých istých účastníkov.

Dvojvýberový t-test

```
> pred<-c(77,56,64,60,57,53,72,62,65,66)
> po<-c(88,74,83,68,58,50,67,64,74,60)
> boxplot(pred,po) #pozriem si graficky
> t.test(pred,po,var.equal=TRUE,alt="less")
        Two Sample t-test
data:  pred and po
t = -1.2484, df = 18, p-value = 0.1139
alternative hypothesis:
true difference in means is less than 0
95 percent confidence interval:
     -Inf 2.100925
sample estimates:
mean of x mean of y
     63.2      68.6
```

Párový t-test

```
> t.test(pred,po,paired=TRUE,alt="less")

        Paired t-test
data:  pred and po
t = -1.8904, df = 9, p-value = 0.04564
alternative hypothesis:
true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1635821
sample estimates:
mean of the differences
                   -5.4
>
```

Párový Wilcoxonov test

Vzhľadom na to, že rozdelenie získaných bodov nie je známe, a náš výber má malý rozsah, je vhodnejšie použiť Wilcoxonov párový test.

```
> wilcox.test(pred,po,paired=TRUE,alt="less")

        Wilcoxon signed rank test

data:  pred and po
V = 12, p-value = 0.06543
alternative hypothesis:
true location shift is less than 0
>
```

χ^2 test dobrej zhody

Účelom testu je overiť, či má náhodná premenná určité predpokladané rozdelenie.

Ako postupujeme:

- obor hondôt náhodnej premennej rozdelíme na k disjunktných častí,
- na základe predpokladaného rozdelenie vypočítame teoretickú pravdepodobnosť p_i, že náhodná premenná nadobudne hodnotu patriacu do i-tej časti, $i = 1, \ldots, k$,
- vykonáme N pokusov, zaznamenajú sa početnosti X_i, koľkokrát hodnota padla do i-tej časti, $i = 1, \ldots, k$,
- porovnajú sa odchýlky empirických početností X_i a teoretických početností Np_i.

χ^2 test dobrej zhody

Pre posúdenie významnosti odchyliek teoretických a empirických početností vypočítame hodnotu testovacej štatistiky

$$\chi^2 = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$$

Ak má náhodná premenná predpokladané rozdelenie, tak testovacia štatistika má rozdelenie chí kvadrát s $(k - s - 1)$ stupňami voľnosti, pričom hodnota s udáva počet neznámych parametrov rozdelenia, ktoré bolo potrebné odhadnúť.

Poznámka: test je možné použiť pri splnení predpokladu, že pre všetky hodnoty je $Np_i > 5$.

χ^2 test dobrej zhody

Pre posúdenie zhody s rovnomerným rozdelením nám stačí jednorozmerný objekt.

tetsujeme teda hypotézu:

H_0: „Všetky kategórie sú rovnako pravdepodobné."

proti alternatíve

H_A: „Všetky kategórie nie sú rovnako pravdepodobné."

Použijeme funkciu `chisq.test()` s jediným argumentom – tabuľkou údajov.

χ^2 test dobrej zhody

Príklad

Bolo vykonaných 300 hodov kockou, s týmito výsledkami:

Počet bodov	1	2	3	4	5	6
Počet padnutí	52	45	48	40	55	60

Chceme posúdiť, či kocka nie je falošná, t.j.j či niektoré čísla nepadajú častejšie.

χ^2 test dobrej zhody

Riešenie

Najskôr musíme zadať namerané hodnoty:

```
> x<-c(52,45,48,40,55,60)
```

Potom vykonáme test:

```
> chisq.test(x)
        Chi-squared test for given probabilities
data:  x
X-squared = 5.16, df = 5, p-value = 0.3967
```

χ^2 test dobrej zhody

Príklad

Na MS 2018 vo futbale bolo odohraných 64 zápasov, v riadnom hracom čase padlo 169 gólov, čo predstavuje 2.64 góla na zápas. Počty gólov dosiahnuté v jednotlivých zápasoch boli takéto:

Počet gólov	0	1	2	3	4	5	6+
Počet zápasov	1	15	18	18	5	2	5

Chceme overiť, či počet gólov dosiahnutých v zápase sa riadi Poissonovým rozdelením.

χ^2 test dobrej zhody

Riešenie

Najskôr definujeme vektor počtov zápasov podľa strelenych golov a potom vygenerujeme vektor poissonovských pravdepodobností.

```
> zapasy<-c(1,15,18,18,5,2,5)
> pnosti<-dpois(strelene[1:6],2.64)
> pnosti[7]<-1-sum(pnosti)
```

Potom vykonáme test:

```
> chisq.test(x=zapasy,p=pnosti)
        Chi-squared test for given probabilities
data:  zapasy
X-squared = 9.4024, df = 6, p-value = 0.1522
```

χ^2 test dobrej zhody

Test nezávislosti

V tomto prípade testujeme nulovú hypotézu

H_0: „Náhodné premenné X a Y sú nezávislé."

Proti alternatíve

H_A: „Náhodné premenné X a Y sú závislé."

χ^2 test dobrej zhody

Test nezávislosti

K dispozícii je náhodný výber, ktorého výsledky sú zoradené do kontingenčnej tabuľky.

	1	2	...	s	$\sum$
1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	n

Označíme n_{ij} počet prípadov, koľkokrát nastala situácia $X = i$ a $Y = j$.
Ďalej definujeme $n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$ a $n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$
Celkový rozsah $n = \sum_{i=1}^{s} \sum_{j=1}^{r} n_{ij}$.

χ^2 test dobrej zhody

Test nezávislosti

K dispozícii je náhodný výber, ktorého výsledky sú zoradené do kontingenčnej tabuľky.

	1	2	...	s	$\sum$
1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
$\sum$	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	n

Určíme teoretické početnosti $e_{ij} = \frac{n_{i\cdot} \cdot n_{\cdot j}}{n}$.
Štatistika $\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$ má rozdelenie chí-kvadrát s $(s-1)(r-1)$ stupňami voľnosti.

χ^2 test dobrej zhody

Test nezávislosti - príklad

V skupine 65 odôb bola zaznamenávaná farba vlasov a farba očí. Výsledky sú v zhrnuté v tabuľke:

	Blond	Hnedá	Čierna
Svetlé	12	8	6
Tmavé	2	25	12

Chceme overiť, či farba očí a vlasov sú nezávislé znaky.

χ^2 test dobrej zhody

Test nezávislosti - riešenie

Najskôr zadáme tabuľku údajov v podobe matice

```
> tabulka=matrix(c(12,2,8,25,6,12),ncol=3)
```

Potom vykonáme test pomocou funkcie `chisq.test()`

```
> chisq.test(tabulka)
        Pearson's Chi-squared test
data:  tabulka
X-squared = 15.938, df = 2, p-value = 0.000346
```

Kolmogorovov – Smirnov test

Neparametrický test, vyšetruje zhodu empirického rozdelenia s teoretickým resp. zhodu dvoch empirických rozdelení.

Jednovýberový test testuje hypotézu:

H_0: „Náhodná premenná má predpokladané (teoretické) rozdelenie.“

proti alternatíve

H_A: „Náhodná premenná nemá predpokladané (teoretické) rozdelenie.“

Testovacia štatistika $D = \sup_x |F_n(x) - F(x)|$

Kolmogorovov – Smirnovov test

Neparametrický test, vyšetruje zhodu empirického rozdelenia s teoretickým resp. zhodu dvoch empirických rozdelení.

Dvojvýberový test testuje hypotézu:

H_0: „Dve jednorozmerné náhodné premenné pochádzajú z rovnakého pravdepodobnostného rozdelenia."

proti alternatíve

H_A: „Dve jednorozmerné náhodné premenné nepochádzajú z rovnakého pravdepodobnostného rozdelenia."

Testovacia štatistika $D = \sup_x \left| F_{1n_1}(x) - F_{2n_1}(x) \right|$

Kolmogorovov – Smirnov test

Príklad

Opakujeme test, či počet gólov strelených na MS2018 v jednom zápase má Poissonovo rozdelenie s parametrom $\lambda = 2.64$.

Test vykonáme pomocou funkcie `ks.test()`, arguenty sú náhdný výber, typ rozdelenia a jeho parametre.

```
> ks.test(zapasy,"ppois",2.64)
        One-sample Kolmogorov-Smirnov test
data:  zapasy
D = 0.66225, p-value = 0.004309
alternative hypothesis: two-sided
```

Kolmogorovov – Smirnovov test

Príklad

Na MS2014 boli počty zápasov, podľa gólov dosiahnutých v riadnom hracom čase takéto:

Počet gólov	0	1	2	3	4	5	6+
Počet zápasov	9	10	10	18	9	4	4

Chceme zistiť, či počty góĺov v rokoch 2014 a 2018 pochádzajú z rovnakého rozdelenia pravdepodobnosti.

Kolmogorovov – Smirnov test

Riešenie

Najskôr definujeme premennu zapasy14 pre rok 2014, potom použijeme funkciu `ks.test()`.

```
> zapasy2<-c(9,10,10,18,9,4,4)
> ks.test(zapasy,zapasy2)
        Two-sample Kolmogorov-Smirnov test
data:  zapasy and zapasy2
D = 0.28571, p-value = 0.9375
alternative hypothesis: two-sided
```