Statistinės duomenų analizės praktinės užduotys

2017

7. **Dažnių lentelės.** Atitinkami R pavyzdžiai

(a) *Hipotezės apie polinominio skirstinio parametrų reikšmes tikrinimas.* (ČM III.5.1) Atliktas tyrimas, kurio tikslas - nustatyti, kokios spalvos automobiliai populiariausi. Atsitiktinai apklausus 200 potencialių pirkėjų, gauti rezultatai, kurie pateikiami lentelėje.

Spalva	Raudona	Geltona	Mėlyna	Žalia	Ruda
Dažnis	39	65	46	37	13

Patikrinkite hipotezę, kad pirkėjai nevienodai vertina automobilių spalvas.

```r
#   Naudojame visur chi-squared (asimptotiniai reiks), bet jei norėtumėm tiksliai
#   tai pabandome su Fisheriu (nors tai nėra būtina)
daznis <- c(39, 65, 46, 37, 13)

t <- chisq.test(daznis)
pv <- t$p.value
t
```

```
##
##  Chi-squared test for given probabilities
##
## data:  daznis
## X-squared = 35, df = 4, p-value = 4.645e-07
```

Kadangi visos kategorijos lygios, tai tikimybinės lentelės nesudarome. $pv = 4.6453484 \times 10^{-7} < \alpha = 0.05$, todėl hipotezę, kad skirtumai tarp kvadratų $(P(spalva) - E(P(Hipotezesspalva)))^2$ yra lygūs nuliui, atmetama. Pirkėjai spalvas vertina nevienodai.

(b) *Hipotezės apie polinominio skirstinio parametrų reikšmes tikrinimas.* (ČM III.5.2) Rinkos analitikas mano, kad A, B, C ir D rūšies pastos vartotojų dalis yra atitinkamai 0.30, 0.60, 0.08, 0.02. Atsitiktinai apklausus 600 žmonių, kokią pastą jie vartoja, gatui rezultatai, kurie pateikiami lentelėje

Rūšis	A	B	C	D
Dažnis	192	342	44	22

Ar šie duomenys leidžia suabejoti rinkos analitiko teiginiu?

$$H : p_1 = p_1^0, p_2 = p_2^0, p_3 = p_3^0, p_4 = p_4^0$$

```r
daznis <- c(192, 342, 44, 22)
tikimybes <- c(0.30, 0.60, 0.08, 0.02)

t <- chisq.test(daznis, p=tikimybes) #cia jau turime tikimybes
pv <- t$p.value
t
```

```
##
```

```
##  Chi-squared test for given probabilities
##
## data:  daznis
## X-squared = 10.367, df = 3, p-value = 0.01569
```

$pv = 0.0156932 < \alpha = 0.05$, todėl hipotezę, kad skirtumai tarp kvadratų, yra lygus nuliui, atmetama. Duomenys leidžia teigti, jog analitikas neteisus.

(c) *Požymių nepriklausomumo tikrinimas.* Lentelėje pateikti skaičiai sutukuotinių, sugrupuotų pagal vaikų skaičių (požymis A) ir metines pajamas (požymis B). Rekia patikrinti hipotezę dėl požymių A ir B nepriklausomumo.

	0 – 1	1 – 2	2 – 3	3	Σ
0	2161	3577	2184	1635	9557
1	2755	5081	2222	1052	11110
2	936	1753	640	306	3635
3	325	419	96	38	878
4	39	98	31	14	182
Σ	6216	10928	5173	3046	25362

```r
data <-  matrix(
    c(2161, 2755, 936, 325, 39,
      3577, 5081, 1753, 419, 98,
      2184, 2222, 640, 96, 31,
      1635, 1052, 306, 38, 14),
    5,
    4,
    dimnames=list(
                   vaiku.sk=c("0", "1", "2", "3","4"),
                   pajamos=c("0-1000", "1000-2000", "2000-3000", "3000+")
                 )
  )

t <- chisq.test(data)
pv <- t$p.value
t
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 628.27, df = 12, p-value < 2.2e-16
```

$pv = 9.6787313 \times 10^{-127} < \alpha = 0.05$, todėl hipotezę, kad $P(Vaiku.sk|Pajamos) = P(Vaiku.sk)\dot{P}(Pajamos)$ arba apie požymių priklausomumą, atmetame.

(d) *Požymių nepriklausomumo tikrinimas* (ČM III.5.4) Buvo tirta, ar užimamos pareigos ir pasitenkinimas darbu yra tarpusavy susiję dalykai. Atsitiktinai apklausus 800 aukštųjų mokyklų dėstytojų, buvo gauti tokie rezultatai:

	Asistentas	Lektorius	Docentas	Profesorius
Patenkintas	40	60	52	63
Neturi nuomonės	78	87	82	88
Nepatenkintas	57	63	66	64

Patikrinkite hipotezę apie pareigų ir pasitenkinimo darbu priklausomybę.

```r
dnames <- list(
        Pasitenkinimas = c("Patenkintas", "Neturi nuomones", "Nepatenkintas"),
        Pareigos = c("Asistentas", "Lektorius", "Docentas", "Profesorius")
      )
Job <- matrix(
        c(40, 78, 57,
          60, 87, 63,
          52, 82, 66,
          63, 88, 64),
        3,
        4,
        dimnames=dnames
      )

fisher.test(Job, workspace=2e7)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  Job
## p-value = 0.8383
## alternative hypothesis: two.sided
```

```r
chisq.test(Job)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Job
## X-squared = 2.7506, df = 6, p-value = 0.8394
```

$pv > \alpha = 0.05$, todel hipotezės, kad $P(Pareigos\|Pasitenkinimas) = P(Pareigos)\dot{P}(Pasitenkinimas)$ arba apie pareigų ir pasitenkinimo darbu priklausomumą negalime atmesti.

(e) *Homogeniškumo hipotezės tikrinimas.* Viename sraute iš 300 stojančiųjų pažymius "nepatenkinamai", "patenkinimai", "gerai" ir "labai gerai" gavo atitinkamai 33, 43, 80 ir 144; kito srauto stojantieji atitinkamai 39, 35, 72, ir 154. Ar galima laikyti, kad abiejų srautų stojantieji pasiruošę vienodai?

```r
dnames <- list(
        Ivertinimas = c("Nepatenkinamai", "Patenkinamai", "Gerai", "Labai gerai"),
        Srautas = c("Pirmas", "Antras")
      )

rezultatai <- matrix(
        c(33, 43, 80, 144,
          39, 35, 72, 154),
        4,
        2,
        dimnames=dnames
      )

fisher.test(rezultatai, workspace=2e7)
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  rezultatai
## p-value = 0.5541
## alternative hypothesis: two.sided
```

```
chisq.test(rezultatai)
```

```
##
##  Pearson's Chi-squared test
##
## data:  rezultatai
## X-squared = 2.0771, df = 3, p-value = 0.5566
```

$pv > \alpha = 0.05$, todėl hipotezės, kad $P(Ivertinimas \| Srautas = 1) = P(Ivertinimas \| Srautas = 1) = P(Ivertinimas)$ arba apie rezultatų priklausymą nuo imties, atmesti negalime.

(f) *Homogeniškumo hipotezės tikrinimas* (ČM III.5.6) Sveikatos apsaugos ministerija tyrė, ar įvairių profesijų žmonių alkoholio vartojimo įpročiai buvo tokie pat. Atsitiktinai apklausus 200 mokytojų, 300 teisininkų ir 400 gydytojų, buvo gauti tokie rezultatai:

	Mokytojai	Teisininkai	Gydytojai
Mažai	100	50	100
Vidutiniškai	50	150	200
Daug	50	100	200

Ar galima teigti, kad šių trijų profesijų atstovų alkoholio vartojimo įpročiai tokie pat?

```
dnames <- list(
        Alkoholis=c("Mazai", "Daug", "Vidutiniskai"),
        Profesija=c("Mokytojas", "Teisininkas", "Daktaras")
      )

sotke <- matrix(
          c(100,50,100, 50,150,200, 50,100,200),
          3,
          3,
          dimnames=dnames
        )

fisher.test(sotke, workspace=2e7)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  sotke
## p-value < 2.2e-16
## alternative hypothesis: two.sided
```

```
chisq.test(sotke)
```

```
##
##  Pearson's Chi-squared test
##
## data:  sotke
## X-squared = 91.607, df = 4, p-value < 2.2e-16
```

Abiems kriterijams $pv < \alpha = 0.05$, todėl hipotezę, kad (tikimybės analogiškai kaip praėjusiam pavyzdy) arba rezultatų priklausymą nuo imties, atmetame.

Padaryta su R version 3.4.2 (2017-09-28), x86_64-pc-linux-gnu.