Krzysztof Trajkowski

Pakiet CressieReadTest

28 maja 2013

Test Cressie-Read

Do badania tabel kontyngencji bardzo często stosuje się testy niezależności χ^2 Pearsona lub G^2 największej wiarygodności. Istnieje jednak bardzo ciekawa (choć mniej popularna) statystyka D^2 zaproponowana przez Reada i Cressie która ma na celu ujednolicenie zapisu całej rodziny statystyk za pomocą poniższego wzoru:

$$D^2 = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \left[\left(\frac{O_{ij}}{E_{ij}} \right)^{\lambda} - 1 \right]$$

gdzie:

- O_{ij} – empiryczna liczebność i-tego wiersza oraz j-tej kolumny,
- E_{ij} – oczekiwana liczebność i-tego wiersza oraz j-tej kolumny,
- r – liczba wierszy,
- c – liczba kolumn.

Zwróćmy uwagę, że wyrażenie $\lambda(\lambda+1)$ musi być różne od zera. A więc parametr λ nie może być równy 0 lub -1. Cressie oraz Read sugerują, aby wartość parametru λ była równa $\frac{2}{3}$ jako kompromis między statystyką χ^2 Pearsona ($\lambda = 1$) i G^2 wskaźnika wiarygodności ($\lambda \to 0$).

```r
m = matrix(c(25, 15, 23, 56), 2, 2)
library(CressieReadTest)
cr.test(m)
```

```
##
##  D-squared Cressie-Read test
##
## data:  m
## D = 12.2532, lambda = 0.6667, df = 1.0000, p-value = 0.0004645
```

Dobierając odpowiednią wartość parametru λ możemy uzyskać wyniki dla kilku różnych testów niezależności opartych na statystyce χ^2. Np. statystykę χ^2 Pearsona otrzymamy gdy $\lambda = 1$, z kolei statystykę Neymana dla $\lambda = -2$ która jest modyfikacją testu χ^2 Pearsona. Statystykę testu G^2 największej wiarygodności uzyskamy dla parametru $\lambda \to 0$, a jego modyfikację Kullback-Leibler gdy $\lambda \to -1$. Natomiast rozwiązanie zaproponowane przez Freemana i Tukeya otrzymamy dla $\lambda = -\frac{1}{2}$. Poniżej przykłady z wykorzystaniem tych testów.

Statystyka chi-kwadrat Pearsona (Pearson chi-squared statistic)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

```r
cgf.test(m, test = "p")
```

```
##
##  X-squared Pearson test
##
## data:  m
## P = 12.3, df = 1.0, p-value = 0.0004532
```

```r
cr.test(m, lambda = 1)
```

```
##
##  D-squared Cressie-Read test (Pearson)
##
## data:  m
## D = 12.3, lambda = 1.0, df = 1.0, p-value = 0.0004532
```

Modyfikacja Neyman's statystyki χ^2:

$$N = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{O_{ij}}$$

```r
cgf.test(m, test = "n")
```

```
##
##  X-squared Neyman's test
##
## data:  m
## N = 13.2, df = 1.0, p-value = 0.0002793
```

```r
cr.test(m, lambda = -2)
```

```
##
##  D-squared Cressie-Read test (Neyman's)
##
## data:  m
## D = 13.2, lambda = -2.0, df = 1.0, p-value = 0.0002793
```

Inna często spotykana statystyka to wskaźnik wiarygodności (log likelihood ratio statistic)

$$G^2 = 2 \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \ln \left(\frac{O_{ij}}{E_{ij}} \right)$$

```r
cgf.test(m, test = "g")
```

```
##
##  G-squared Likelihood Ratio test
##
## data:  m
## G = 12.27, df = 1.00, p-value = 0.0004603
```

```r
cr.test(m, lambda = 1e-05)
```

```
##
##  D-squared Cressie-Read test (G-squared)
##
## data:  m
## D = 12.27, lambda = 0.00, df = 1.00, p-value = 0.0004603
```

Modyfikacja Kullback-Leibler statystyki G^2:

$$KL = 2\sum_{i=1}^{r}\sum_{j=1}^{c} E_{ij}\ln\left(\frac{E_{ij}}{O_{ij}}\right)$$

```
cgf.test(m, test = "kl")

##
##  G-squared Kullback-Leibler test
##
## data:  m
## KL = 12.57, df = 1.00, p-value = 0.0003929

cr.test(m, lambda = -0.99999)

##
##  D-squared Cressie-Read test (Kullback-Leibler)
##
## data:  m
## D = 12.57, lambda = -1.00, df = 1.00, p-value = 0.0003929
```

Poniżej statystyka zaproponowana przez Freemana i Tukeya:

$$FT = 4\sum_{i=1}^{r}\sum_{j=1}^{c}\left(\sqrt{O_{ij}} - \sqrt{E_{ij}}\right)^2$$

```
cgf.test(m, test = "ft")

##
##  F-squared Freeman-Tukey's test
##
## data:  m
## FT = 12.38, df = 1.00, p-value = 0.0004347

cr.test(m, lambda = -0.5)

##
##  D-squared Cressie-Read test (Freeman-Tukey's)
##
## data:  m
## D = 12.38, lambda = -0.50, df = 1.00, p-value = 0.0004347
```

Na bazie statystyki χ^2 można obliczyć kilka współczynników, które określają siłę związku badanych zmiennych.

- współczynnik Yule'a – ma zastosowanie dla tabel o wymiarach 2×2 oraz $\phi \in \langle -1; 1\rangle$:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

- współczynnik Pearsona – ma zastosowanie dla tabel o wymiarach $r \times c$ oraz $C \in \left\langle 0; \sqrt{\frac{min(r,c)-1}{min(r,c)}} \right\rangle$:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- współczynnik Cramera – nie wskazuje kierunku korelacji oraz $V \in \langle 0; 1 \rangle$:

$$V = \sqrt{\frac{\chi^2}{n(min(r,c) - 1)}}$$

- współczynnik Czupurowa – nie wskazuje kierunku korelacji oraz $T \in \langle 0; 1 \rangle$:

$$T = \sqrt{\frac{\chi^2}{n\sqrt{(r-1)(c-1)}}}$$

Poniżej są przedstawione obliczenia dla wszystkich omówionych testów oraz współczynniki korelacji:

```
allcr.test(m)

## $test
##                           lambda Statistic df     p-val
## Pearson's Chi-squared   1.0000     12.30  1 0.0004532
## Log likelihood ratio    0.0000     12.27  1 0.0004603
## * Williams correction   0.0000     12.08  1 0.0005086
## Cressie-Read            0.6667     12.25  1 0.0004645
## Freeman-Tukey's        -0.5000     12.38  1 0.0004347
## Neyman's               -2.0000     13.20  1 0.0002793
## Kullback-Leibler       -1.0000     12.57  1 0.0003929
##
## $coefficient
##               Y-Yule'a C-Pearson V-Cramer T-Czupurow
## Coefficient:    0.3215    0.3061   0.3215     0.3215
```

Przedstawione powyżej formuły matematyczne (testy niezależności) są także wykorzystywane do badania zgodności danych liczbowych z określnonym rozkładem np. jednostajnym. Poniżej przykłady dla dwóch rozkładów jednostajnych.

Rozkład jednostajny–dyskretny:

```
set.seed(8746)
s = sample(0:6, 150, T)
table(s)

## s
##  0  1  2  3  4  5  6
## 29 14 27 21 20 20 19

w = as.numeric(table(s))
cr.gof(w)
```

```
##
##  D-squared Cressie-Read test for given probabilities
##
## data:  w
## D = 7.1584, lambda = 0.6667, df = 6.0000, p-value = 0.3064
```

```r
cr.gof(w, lambda = 1)
```

```
##
##  D-squared Cressie-Read test (Pearson) for given probabilities
##
## data:  w
## D = 7.173, lambda = 1.000, df = 6.000, p-value = 0.3051
```

Rozkład jednostajny–ciągły:

```r
set.seed(3254)
s = runif(150)
p = seq(0, 1, 0.2)
table(cut(s, p))
```

```
##
##   (0,0.2] (0.2,0.4] (0.4,0.6] (0.6,0.8]   (0.8,1]
##        21        29        36        30        34
```

```r
w = as.numeric(table(cut(s, p)))
cr.gof(w)
```

```
##
##  D-squared Cressie-Read test for given probabilities
##
## data:  w
## D = 4.5347, lambda = 0.6667, df = 4.0000, p-value = 0.3385
```

```r
cr.gof(w, lambda = 1)
```

```
##
##  D-squared Cressie-Read test (Pearson) for given probabilities
##
## data:  w
## D = 4.467, lambda = 1.000, df = 4.000, p-value = 0.3465
```