Chi squared test

Klinkenberg

12 okt 2017

Inhoud

- chi2 test
  - chi2 test statistic
  - chi2 distribution
  - calculating chi2
  - calculating the model
  - observed model
  - testing for significance
  - fishers exact test
  - yatess correction
  - likelihood ratio
  - standardized residuals
  - effect size

Relation between categorical variables

χ^2 test

χ^2 test

A '''chi-squared test''', also written as χ^2 test, is any statistical hypothesis test wherein the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.

Chi-squared tests are often constructed from a Lack-of-fit sum of squares#Sums of squares|sum of squared errors, or through the Variance Distribution of the sample variance|sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Source: wikipedia

χ^2 test statistic

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}}$$

Contingency table

$$\text{observed}_{ij} = \begin{pmatrix} o_{11} & o_{12} & \cdots & o_{1j} \\ o_{21} & o_{22} & \cdots & o_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ o_{i1} & o_{i2} & \cdots & o_{ij} \end{pmatrix} \text{model}_{ij} = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1j} \\ m_{21} & m_{22} & \cdots & m_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{i1} & m_{i2} & \cdots & m_{ij} \end{pmatrix}$$

χ^2 distribution

The χ^2 distribution describes the test statistic under the assumption of H_0, given the degrees of freedom.

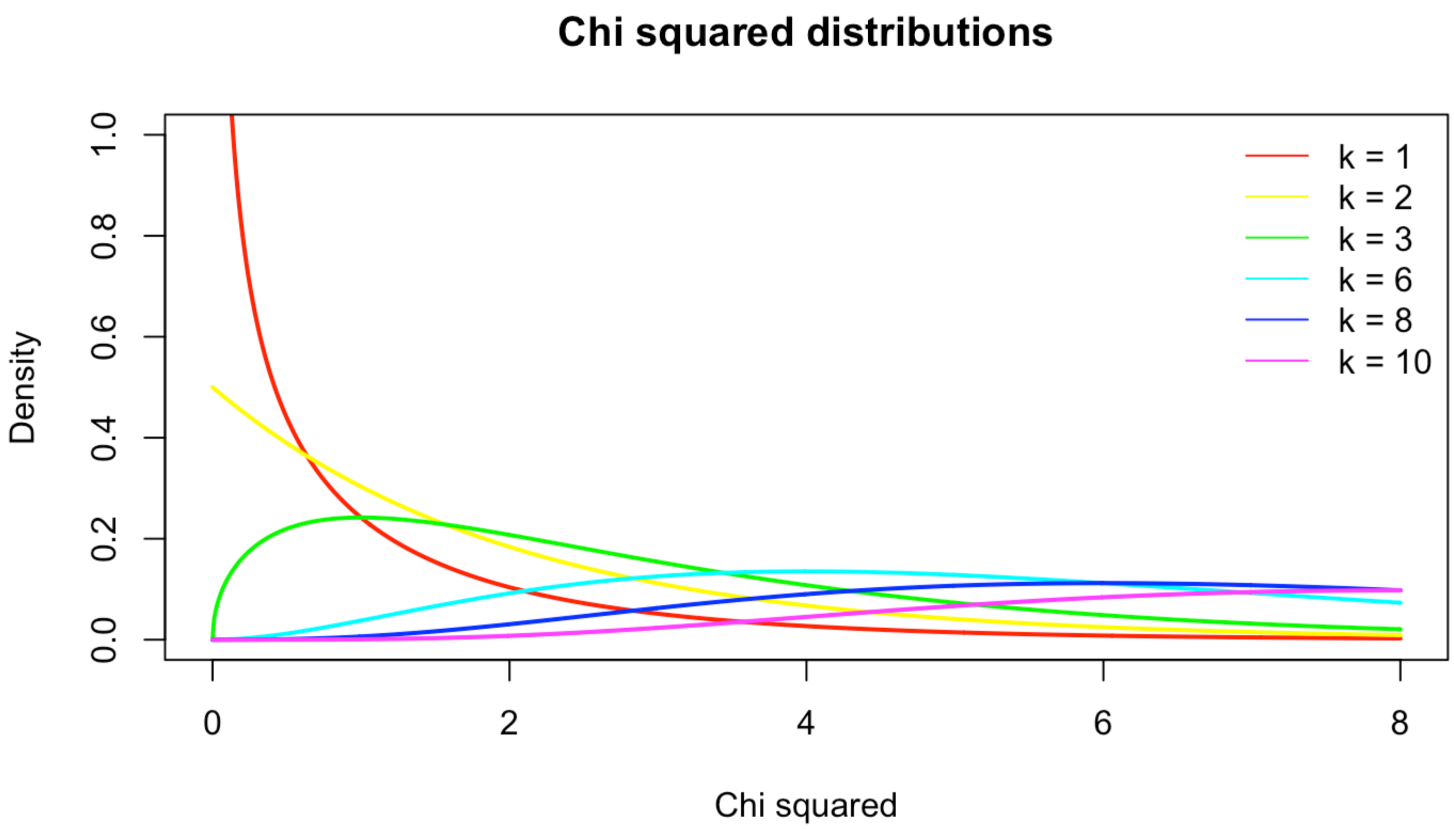
$df = (r-1)(c-1)$ where r is the number of rows and c the amount of columns.

```
chi = seq(0,8,.01)
df  = c(1,2,3,6,8,10)
col = rainbow(n = length(df))

plot( chi, dchisq(chi, df[1]), lwd = 2, col = col[1], type="l",
      main = "Chi squared distributions",
      ylab = "Density",
      ylim = c(0,1),
      xlab = "Chi squared")

lines(chi, dchisq(chi, df[2]), lwd = 2, col = col[2], type="l")
lines(chi, dchisq(chi, df[3]), lwd = 2, col = col[3], type="l")
lines(chi, dchisq(chi, df[4]), lwd = 2, col = col[4], type="l")
lines(chi, dchisq(chi, df[5]), lwd = 2, col = col[5], type="l")
lines(chi, dchisq(chi, df[6]), lwd = 2, col = col[6], type="l")

legend("topright", legend = paste("k =",df), col = col, lty = 1, bty = "n")
```


Chi squared distributions
Density
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
Chi squared
k = 1
k = 2
k = 3
k = 6
k = 8
k = 10

Example

Experiment

http://goo.gl/faj76B

Data

Calculating χ^2

```
observed <- table(data[,c('fluiten','sekse')])
observed
```

```
##          sekse
## fluiten Man Vrouw
##      Ja   17    26
##      Nee   1    12
```

$$\text{observed}_{ij} = \begin{pmatrix} 17 & 26 \\ 1 & 12 \end{pmatrix}$$

Calculating the model

$$\text{model}_{ij} = E_{ij} = \frac{\text{row total}_i \times \text{column total}_j}{n}$$

```
n   = sum(observed)
ct1 = colSums(observed)[1]
ct2 = colSums(observed)[2]
rt1 = rowSums(observed)[1]
rt2 = rowSums(observed)[2]

addmargins(observed)
```

```
##        sekse
## fluiten Man Vrouw Sum
##     Ja   17    26  43
##     Nee   1    12  13
##     Sum  18    38  56
```

Calculating the model

$$\text{model}_{ij} = E_{ij} = \frac{\text{row total}_i \times \text{column total}_j}{n}$$

```
model = matrix( c((ct1*rt1)/n,
                  (ct2*rt1)/n,
                  (ct1*rt2)/n,
                  (ct2*rt2)/n),2,2,byrow=T
          )
model
```

```
##           [,1]      [,2]
## [1,] 13.821429 29.178571
## [2,]  4.178571  8.821429
```

$$\text{model}_{ij} = \begin{pmatrix} 13.8214286 & 29.1785714 \\ 4.1785714 & 8.8214286 \end{pmatrix}$$

observed - model

```
observed - model
```

```
##          sekse
## fluiten         Man     Vrouw
##      Ja   3.178571 -3.178571
##      Nee -3.178571  3.178571
```

Calculating χ^2

$$\chi^2 = \sum \frac{(\text{observed}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}}$$

```
# Calculate chi squared
chi.squared <- sum((observed - model)^2/model)
chi.squared
```

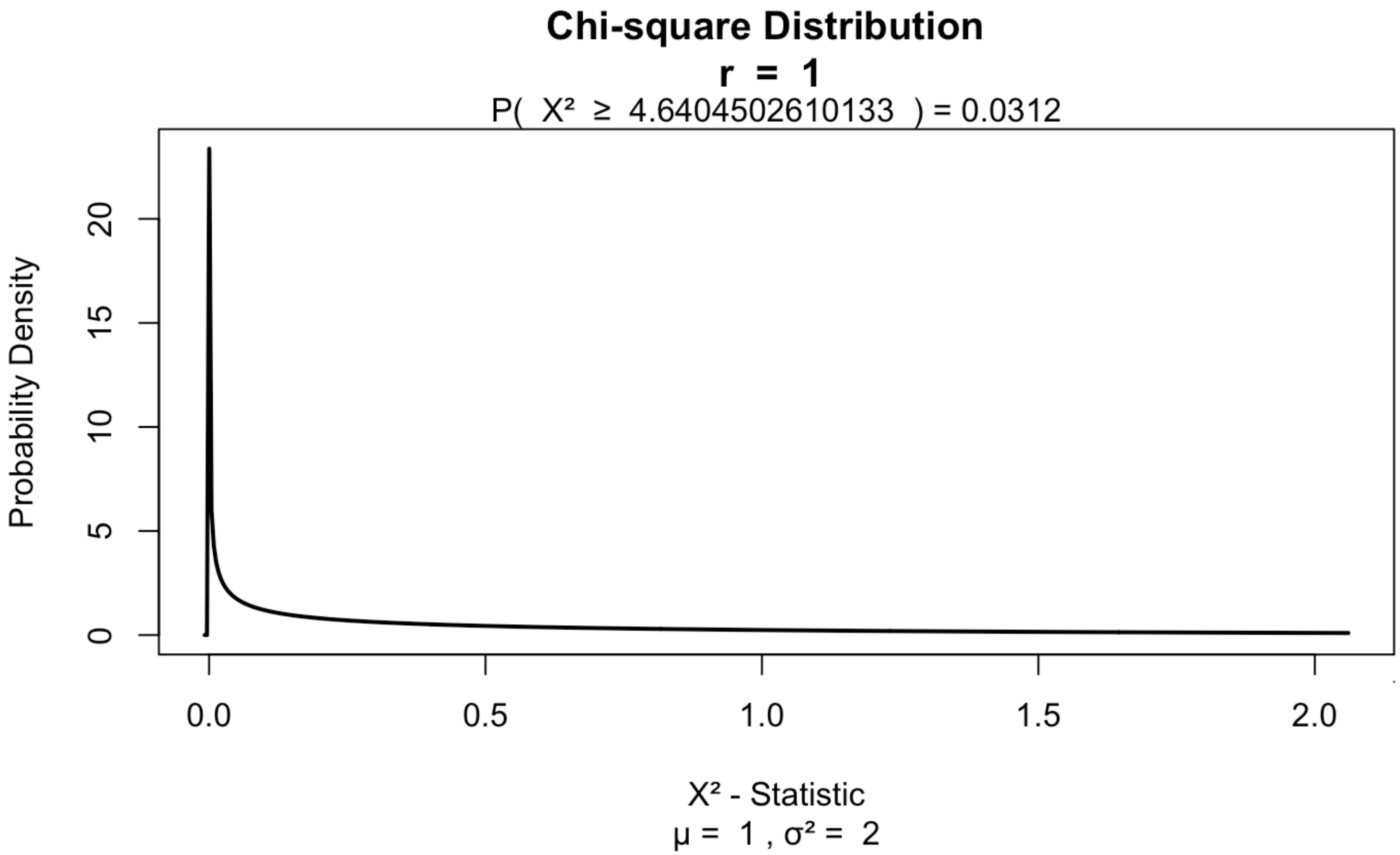
```
## [1] 4.64045
```

Testing for significance

$$df = (r - 1)(c - 1)$$

```
df = (2 - 1) * ( 2 - 1)

library('visualize')
visualize.chisq(chi.squared,df,section='upper')
```


Chi-square Distribution
r = 1
P(X² ≥ 4.6404502610133) = 0.0312
Probability Density
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
X² - Statistic
μ = 1 , σ² = 2

Fisher's exact test

Calculates axact χ^2 for small samples.

- Cell size < 5

Yates's correction

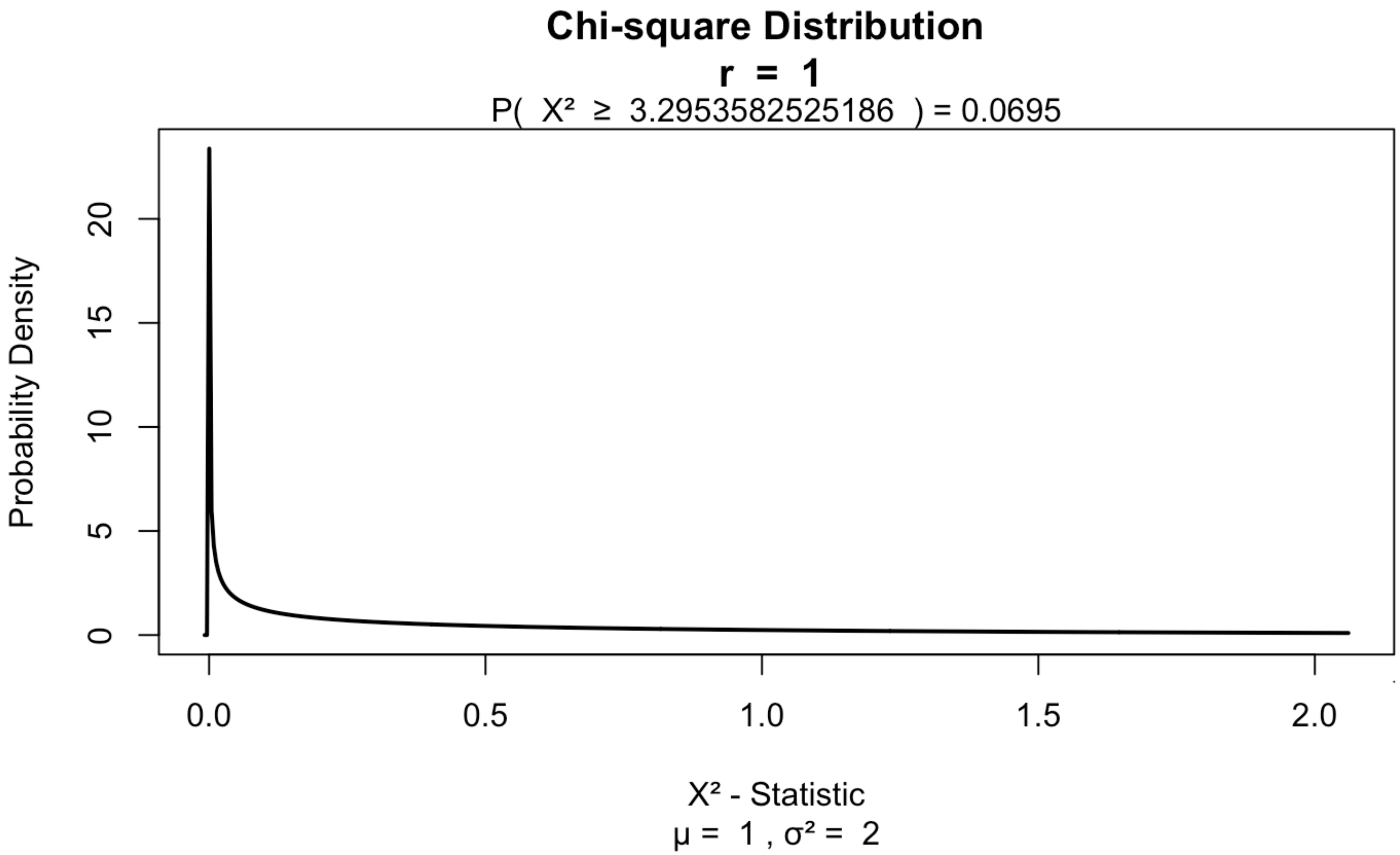
For 2 x 2 contingency tables.

$$\chi^2 = \sum \frac{(|\text{observed}_{ij} - \text{model}_{ij}| - .5)^2}{\text{model}_{ij}}$$

```
# Calculate Yates's corrected chi squared
chi.squared.yates <- sum((abs(observed - model) - .5)^2/model)
chi.squared.yates
```

```
## [1] 3.295358
```

```
visualize.chisq(chi.squared.yates,df,section='upper')
```


Chi-square Distribution
r = 1
P(X² ≥ 3.2953582525186) = 0.0695
Probability Density
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
X² - Statistic
μ = 1 , σ² = 2

Likelihood ratio

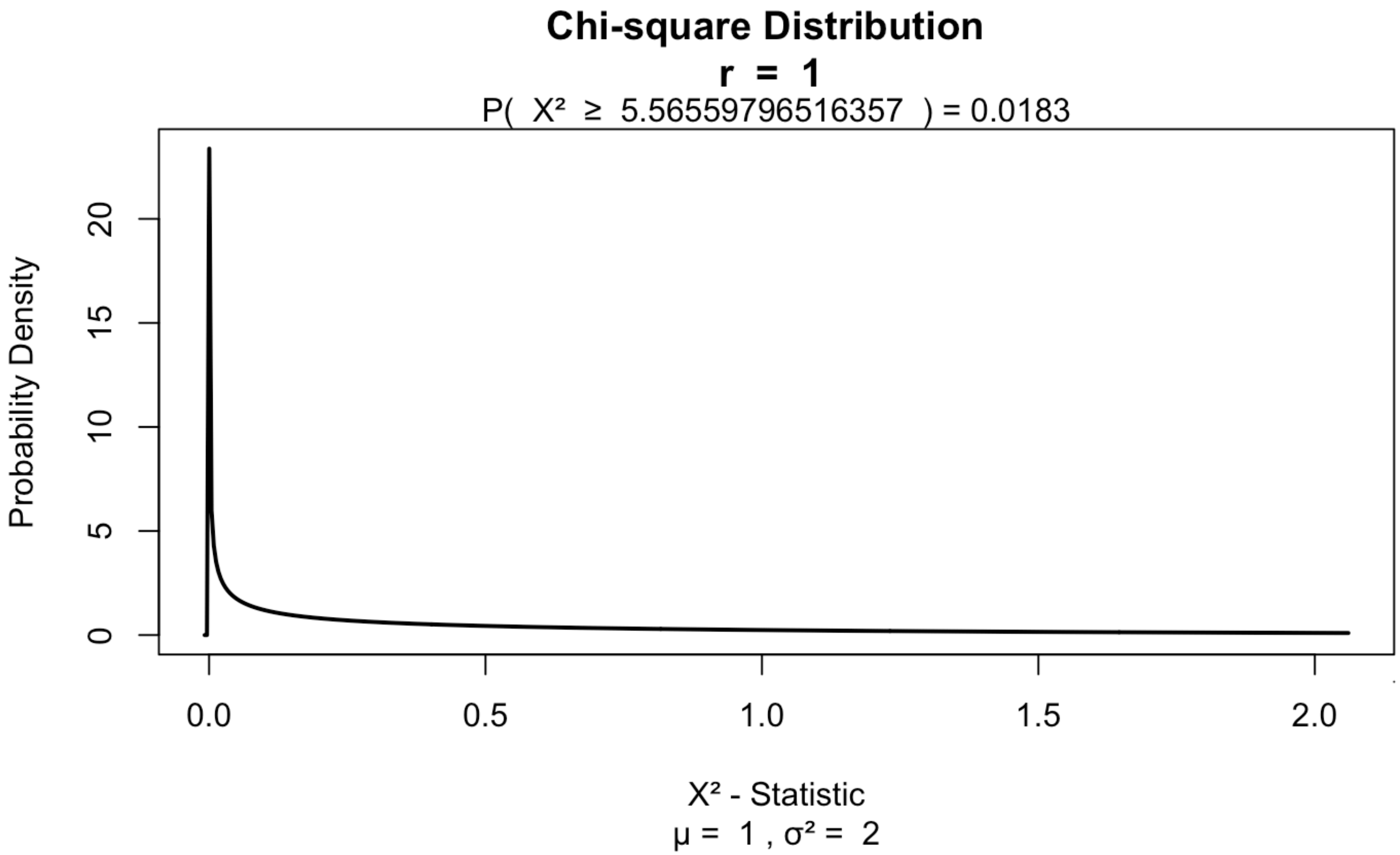
Alternatieve to Pearson's χ^2.

$$L\chi^2 = 2 \sum \text{observed}_{ij} ln \left(\frac{\text{observed}_{ij}}{\text{model}_{ij}} \right)$$

```
# ln is log
lx2 = 2 * sum(observed * log(observed / model) ); lx2
```

```
## [1] 5.565598
```

```
visualize.chisq(lx2,df,section='upper')
```


Chi-square Distribution
r = 1
P(X² ≥ 5.56559796516357) = 0.0183
Probability Density
0
5
10
15
20
0.0
0.5
1.0
1.5
2.0
X² - Statistic
μ = 1 , σ² = 2

Standardized residuals

$$\text{standardized residuals} = \frac{\text{observed}_{ij} - \text{model}_{ij}}{\sqrt{\text{model}_{ij}}}$$

```
(observed - model) / sqrt(model)
```

```
##         sekse
## fluiten        Man      Vrouw
##      Ja  0.8549791 -0.5884370
##      Nee -1.5549558  1.0701940
```

Effect size

Odds ratio based on the observed values

```
odds <- round( observed, 2); odds
```

```
##          sekse
## fluiten Man Vrouw
##      Ja  17    26
##     Nee   1    12
```

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$OR = \frac{a \times d}{b \times c} = \frac{17 \times 12}{26 \times 1} = 7.8461538$$

Odds

```
##         sekse
## fluiten Man Vrouw
##      Ja   17    26
##      Nee   1    12
```

The man and women ratio of people that can whisle and the ratio of those who can't whistle

- Can wistle $\text{Odds}_{mf} = \frac{17}{26}$ = 0.6538462
- Can't wistle $\text{Odds}_{mf} = \frac{1}{12}$ = 0.0833333

Odds ratio

Is the ratio of these odds.

$$OR = \frac{\text{wistle}}{\text{can't wistle}} = \frac{0.6538462}{0.0833333} = 7.8461538$$

- @shklinkenberg
- Klinkenberg
- S.Klinkenberg@UvA.nl
- ShKlinkenberg

END