χ²-tests in R

I. Basic uses of chisq.test(): Pearson’s χ²-tests

chisq.test( x, p = <vector-of-probabilities> )

x: a numeric vector or matrix
p: a vector of probabilities of the same length of ‘x’

If ‘x’ is a matrix with at least two rows and columns, it is taken as a two-dimensional contingency table: the entries of ‘x’ must be non-negative integers.

Goodness-of-fit test
x is a vector => ‘x’ is treated as a one-dimensional contingency table

Example:
\[
x <- c(89,37,30,28,2)
p <- c(0.40,0.20,0.20,0.15,0.05)
chisq.test(x, p = p)
\]

II. Examples based on real data

Goodness-of-fit test

The data comes from the word sense disambiguation task in which the senses of the noun line are investigated. The estimated probabilities are relative frequencies observed in the training dataset. The null hypothesis is that in the test dataset the senses have the same distribution. We will check the hypothesis using Pearson’s χ²-test.

1. Data

<table>
<thead>
<tr>
<th>SENSES</th>
<th>estimated probabilities</th>
<th>test set observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>cord</td>
<td>9.2%</td>
<td>37</td>
</tr>
<tr>
<td>division</td>
<td>8.9%</td>
<td>51</td>
</tr>
<tr>
<td>formation</td>
<td>8.1%</td>
<td>52</td>
</tr>
<tr>
<td>phone</td>
<td>10.6%</td>
<td>44</td>
</tr>
<tr>
<td>product</td>
<td>53.5%</td>
<td>268</td>
</tr>
<tr>
<td>text</td>
<td>9.8%</td>
<td>48</td>
</tr>
</tbody>
</table>

> x = c(37, 51, 52, 44, 268, 48)
> p = c(9.2, 8.9, 8.1, 10.6, 53.5, 9.8)
2. The formula for Pearson’s cumulative test statistic

\[ X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \]

3. Computing the statistic in R “by hands”

```r
> O = x
> E = p/100 * sum(x)

# The statistic:
> sum((O-E)*(O-E)/E)
[1] 7.525384

# The critical value of chi-square with df=5 at 95%:
> qchisq(0.95, df=5)
[1] 11.0705
```

4. Conclusion

the critical value > the computed statistic \(\implies\) we cannot reject the hypothesis

that senses are distributed as we estimated

5. The same using chisq.test()

```r
> chisq.test(x, p=p, rescale.p=T)
Chi-squared test for given probabilities
data: x
X-squared = 7.5324, df = 5, p-value = 0.184
```

Test of independence

The data comes from the word sense disambiguation task in which the patterns of the verb *submit* are recognized. We have a set of (selected) features (in the file “submit.fv”) and will test if they are (statistically) independent. For a pair of features we have a null hypothesis that they are independent.

The features:

- the verb in passive-voice ($pVoice$)
- a nominal-like word just before the verb ($nominal_like$)
- word “to” just before the verb ($to$)
- lemma “be” just before the verb ($to_be$)
- an adverbial just after the verb ($adv.3$)
Observing the contingency tables using R:

```r
> data.submit = read.table("submit.fv", header=T)

> table(data.submit$nominal_like, data.submit$to)
  0 1
  0 119 52
  1 79 0

> table(data.submit$pVoice, data.submit$to)
  0 1
  0 153 52
  1 45 0

> table(data.submit$pVoice, data.submit$nominal_like)
  0 1
  0 126 79
  1 45 0

> table(data.submit$pVoice, data.submit$to_be)
  0 1
  0 204 1
  1 2 43
```

So far, the investigated pairs of features were not independent, obviously.

BUT here we are not clear:

```r
> table(data.submit$pVoice, data.submit$adv.3)
  0 1
  0 162 43
  1 24 21
```

So, we need to use the $\chi^2$-test.

1. The observed and the expected frequencies

<table>
<thead>
<tr>
<th></th>
<th>observed</th>
<th>expected (if they are independent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(pVoice = 0, adv.3 = 0)</td>
<td>162</td>
<td>p(pVoice = 0) * p(adv.3 = 0) * N</td>
</tr>
<tr>
<td>(pVoice = 1, adv.3 = 0)</td>
<td>24</td>
<td>p(pVoice = 1) * p(adv.3 = 0) * N</td>
</tr>
<tr>
<td>(pVoice = 0, adv.3 = 1)</td>
<td>43</td>
<td>p(pVoice = 0) * p(adv.3 = 1) * N</td>
</tr>
<tr>
<td>(pVoice = 1, adv.3 = 1)</td>
<td>21</td>
<td>p(pVoice = 1) * p(adv.3 = 1) * N</td>
</tr>
</tbody>
</table>

$E_{0,0} = (162 + 43) \times (162 + 24) / 250 = 152.52$

$E_{1,0} = (162 + 24) \times (24 + 21) / 250 = 33.48$

$E_{0,1} = (162 + 43) \times (43 + 21) / 250 = 52.48$

$E_{1,1} = (43 + 21) \times (24 + 21) / 250 = 11.52$

2. The formula for Pearson’s cumulative test statistic

$$X^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$
3. Computing the statistic in R “by hands”

\[ X^2 = \frac{(162-152.52)^2}{152.52} + \frac{(24-33.48)^2}{33.48} + \frac{(43-52.48)^2}{52.48} + \frac{(21-11.52)^2}{11.52} \]

\[ X^2 = 12.78726 \]

# The critical value of chi-square with df=1 at 95%
# (here df=1 because df=(r-1)*(c-1))
> qchisq(0.95, df=1)
[1] 3.841459

4. Conclusion

the critical value < the computed statistic \(\implies\) we reject the hypothesis that the two features are independent

5. The same using chisq.test()

> x=table(data.submit$pVoice, data.submit$adv.3)
> chisq.test(x)

Pearson’s Chi-squared test with Yates’ continuity correction

data: x
X-squared = 11.474, df = 1, p-value = 0.0007058

# and without the "correction"
> chisq.test(x, simulate.p.value=T)

Pearson’s Chi-squared test with simulated p-value (based on 2000 replicates)

data: x
X-squared = 12.7873, df = NA, p-value = 0.0004998