Theory
Statistical Theory: Chi-Square
Application to the Titanic Dataset

Christoph Traumüller

24.01.2025

Chi-Square Test of Independence

The chi-square test of independence requires the following assumptions:

- Both variables are categorical.
- Observations are independent.
- Expected cell frequencies are sufficiently large (commonly $E_{ij} \geq 5$).

In the Titanic dataset, the variables *Sex* (male, female) and *Survived* (yes, no) are categorical, thus fulfilling the first assumption of the chi-square test of independence. Furthermore, the observations can be assumed to be independent. Finally, the expected cell frequencies are evaluated using the formula

$$E_{ij} = \frac{(\text{row sum}_i)(\text{column sum}_j)}{n},$$

where n denotes the total sample size. All expected cell frequencies are sufficiently large ($E_{ij} \geq 5$), satisfying the third assumption.

```
table(survived)

## survived
##  0  1
## 79 44

table(male)

## male
##  0  1
## 46 77

table(male, survived)
```

```
##     survived
## male  0  1
##    0 13 33
##    1 66 11
```

```
tab <- table(survived, male)
chisq.test(tab)$expected
```

```
##         male
## survived        0        1
##        0 29.54472 49.45528
##        1 16.45528 27.54472
```

The null hypothesis is defined as:

$$H_0 : \mathit{Sex} \perp \mathit{Survived}.$$

Under the null hypothesis, gender and survival status are assumed to be statistically independent. The chi-square test statistic is defined as

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

In the present 2 × 2 case, the chi-square statistic reduces to

$$\chi^2 = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}.$$

The expected frequencies are computed as follows:

$$\begin{aligned} E_{11} &= \frac{46 \cdot 79}{123} = 29.54, \\ E_{12} &= \frac{46 \cdot 44}{123} = 16.46, \\ E_{21} &= \frac{77 \cdot 79}{123} = 49.46, \\ E_{22} &= \frac{77 \cdot 44}{123} = 27.54. \end{aligned}$$

Using the observed frequencies obtained from `table(male, survived)`, the chi-square statistic is calculated as

$$\begin{aligned} \chi^2 &= \frac{273.57}{29.54} + \frac{273.57}{16.46} + \frac{273.57}{49.46} + \frac{273.57}{27.54} \\ &= 9.26 + 16.62 + 5.53 + 9.94 \\ &= 41.35. \end{aligned}$$

For comparison, the chi-square test is also computed using the `chisq.test` function in R:

```
chisq.test(table(male, survived), correct = FALSE)
```

```
##
```

```
##  Pearson's Chi-squared test
##
## data:  table(male, survived)
## X-squared = 41.372, df = 1, p-value = 1.259e-10
```

Under the null hypothesis, the test statistic χ^2 follows a chi-square distribution with

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

In our case, the contingency table is of size 2×2, and therefore $df = (2-1)(2-1) = 1$.
Large values of χ^2 indicate evidence against independence.
The chi-square statistic was computed both manually and using the `chisq.test` function in R. In both cases, the result is identical, yielding $\chi^2 = 41.35$ with one degree of freedom, indicating strong evidence against the null hypothesis of independence.