# Chi-square and ordinal tests

The ASTA team

## Contents



# 1 Contingency tables

## 1.1 A contingency table

- The dataset `popularKids`, we study the **association** between the **factors** `Goals` and `Urban.Rural`:
  - `Urban.Rural`: The students were selected from urban, suburban, and rural schools.
  - `Goals`: The students indicated whether good grades, athletic ability, or popularity was most important to them.
- Based on a sample we make a cross tabulation of the factors and we get a so-called **contingency table** (krydstabel).

```r
popKids <- read.delim("https://asta.math.aau.dk/datasets?file=PopularKids.dat")
library(mosaic)
tab <- tally(~Urban.Rural + Goals, data = popKids, margins = TRUE)
tab
```

```
##             Goals
## Urban.Rural Grades Popular Sports Total
##    Rural        57      50     42   149
##    Suburban     87      42     22   151
##    Urban       103      49     26   178
##    Total       247     141     90   478
```

## 1.2 A conditional distribution

- Another representation of data is the probability distribution of `Goals` for each level of `Urban.Rural`, i.e. the sum in each row of the table is 1 (up to rounding):

```
##             Goals
## Urban.Rural Grades Popular Sports   Sum
##    Rural     0.383   0.336  0.282 1.000
##    Suburban  0.576   0.278  0.146 1.000
##    Urban     0.579   0.275  0.146 1.000
##    Total     0.517   0.295  0.188 1.000
```

- Here we will talk about the **conditional distribution** of `Goals` given `Urban.Rural`.
- An important question could be:
  - Are the goals of the kids different when they come from urban, suburban or rural areas? I.e. are the rows in the table significantly different?
- There is (almost) no difference between urban and suburban, but it looks like rural is different.

## 1.3 Independence

- Recall, that two factors are **independent**, when there is no difference between the population's distributions of one factor given the levels of the other factor.
- Otherwise the factors are said to be **dependent**.
- If we e.g. have the following conditional **population distributions** of `Goals` given `Urban.Rural`:

```
##             Goals
## Urban.Rural Grades Popular Sports
##    Rural       0.5     0.3    0.2
##    Suburban    0.5     0.3    0.2
##    Urban       0.5     0.3    0.2
```

- Then the factors `Goals` and `Urban.Rural` are independent.
- We take a sample and "measure" the factors $F_1$ and $F_2$. E.g. `Goals` and `Urban.Rural` for a random child.
- The hypothesis of interest today is:

$$H_0 : F_1 \text{ and } F_2 \text{ are independent}, \quad H_a : F_1 \text{ and } F_2 \text{ are dependent}.$$

## 1.4 The Chi-squared test for independence

- Our best guess of the distribution of `Goals` is the relative frequencies in the sample:

```r
tab <- tally(~Urban.Rural + Goals, data = popKids)
n <- margin.table(tab)
pctGoals <- round(margin.table(tab, 2) / n, 3)
pctGoals
```

```
## Goals
##  Grades Popular  Sports
##   0.517   0.295   0.188
```

- If we assume independence, then this is also a guess of the conditional distributions of `Goals` given `Urban.Rural`.
- The corresponding expected counts in the sample are then:

```
##            Goals
## Urban.Rural Grades        Popular       Sports        Sum
##    Rural     77.0 (0.517)  44.0 (0.295)  28.1 (0.188) 149.0 (1.000)
##    Suburban  78.0 (0.517)  44.5 (0.295)  28.4 (0.188) 151.0 (1.000)
##    Urban     92.0 (0.517)  52.5 (0.295)  33.5 (0.188) 178.0 (1.000)
##    Sum      247.0 (0.517) 141.0 (0.295)  90.0 (0.188) 478.0 (1.000)
```

## 1.5 Calculation of expected table

```r
pctexptab
```

```
##            Goals
## Urban.Rural Grades        Popular       Sports        Sum
##    Rural     77.0 (0.517)  44.0 (0.295)  28.1 (0.188) 149.0 (1.000)
##    Suburban  78.0 (0.517)  44.5 (0.295)  28.4 (0.188) 151.0 (1.000)
##    Urban     92.0 (0.517)  52.5 (0.295)  33.5 (0.188) 178.0 (1.000)
##    Sum      247.0 (0.517) 141.0 (0.295)  90.0 (0.188) 478.0 (1.000)
```

- We note that
  - The relative frequency for a given column is **column total** divided by **table total**. For example `Grades`, which is $\frac{247}{478} = 0.517$.
  - The expected value in a given cell in the table is then the cell's relative column frequency multiplied by the cell's **row total**. For example `Rural` and `Grades`: $149 \times 0.517 = 77.0$.
- This can be summarized to:
  - The expected value in a cell is the product of the cell's **row total** and **column total** divided by the **table total**

## 1.6 Chi-squared ($\chi^2$) test statistic

- We have an **observed table**:

```r
tab
```

```
##            Goals
## Urban.Rural Grades Popular Sports
##    Rural        57      50     42
##    Suburban     87      42     22
##    Urban       103      49     26
```

- And an **expected table**, if $H_0$ is true:

```
##             Goals
## Urban.Rural Grades Popular Sports  Sum
##    Rural      77.0   44.0    28.1  149.0
##    Suburban  78.0    44.5    28.4  151.0
##    Urban      92.0   52.5    33.5  178.0
##    Sum       247.0  141.0    90.0  478.0
```

- If these tables are "far from each other", then we reject $H_0$. We want to measure the distance via the Chi-squared test statistic:
  - $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$: Sum over all cells in the table
  - $f_o$ is the frequency in a cell in the observed table
  - $f_e$ is the corresponding frequency in the expected table.
- We have:

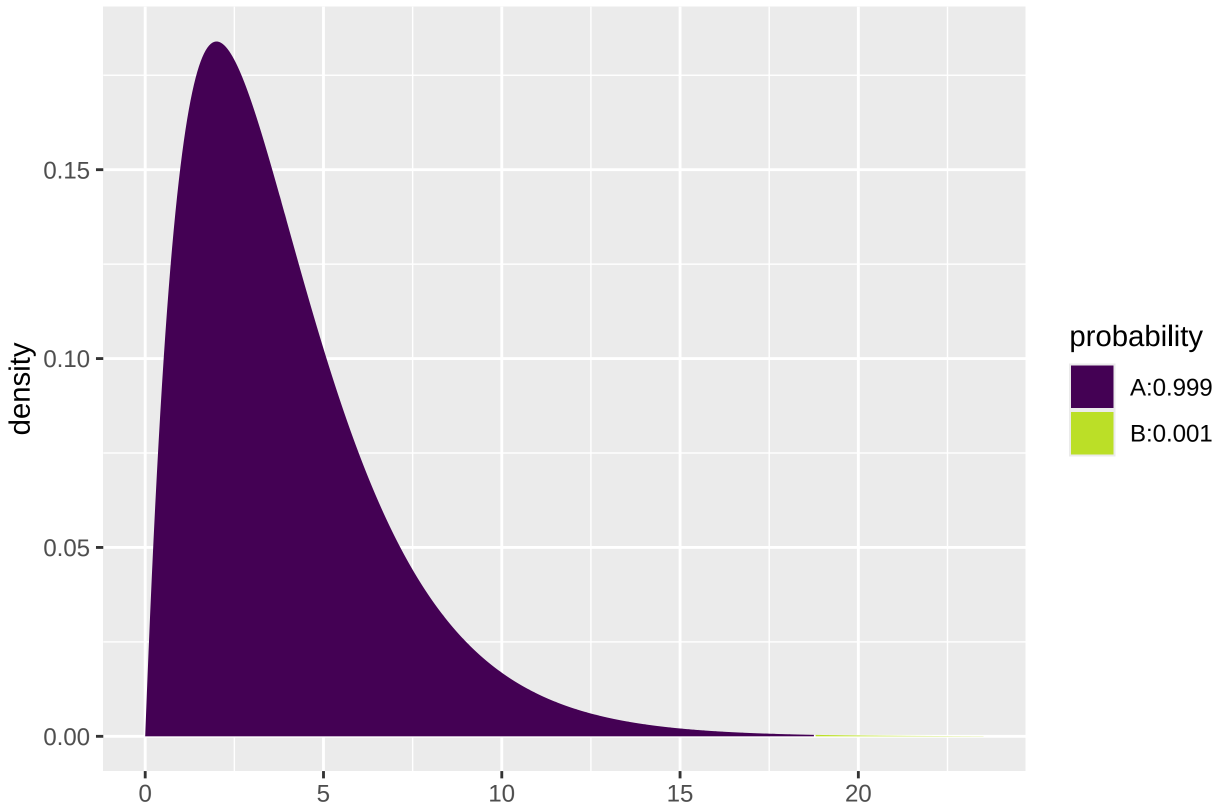
$$X^2_{obs} = \frac{(57-77)^2}{77} + \ldots + \frac{(26-33.5)^2}{33.5} = 18.8$$

- Is this a large distance??

## 1.7 $\chi^2$-test template.

- We want to test the hypothesis $H_0$ of independence in a table with $r$ rows and $c$ columns:
  - We take a sample and calculate $X^2_{obs}$ - the observed value of the test statistic.
  - p-value: Assume $H_0$ is true. What is then the chance of obtaining a larger $X^2$ than $X^2_{obs}$, if we repeat the experiment?
- This can be approximated by the **$\chi^2$-distribution** with $df = (r-1)(c-1)$ degrees of freedom.
- For `Goals` and `Urban.Rural` we have $r = c = 3$, i.e. $df = 4$ and $X^2_{obs} = 18.8$, so the p-value is:

```
1 - pdist("chisq", 18.8, df = 4)
```




```
## [1] 0.0008603303
```

- There is clearly a significant association between `Goals` and `Urban.Rural`.

## 1.8 The function `chisq.test`

- All of the above calculations can be obtained by the function `chisq.test`.

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 18.828, df = 4, p-value = 0.0008497
```

```
testStat$expected
```

```
##            Goals
## Urban.Rural   Grades  Popular   Sports
##    Rural    76.99372 43.95188 28.05439
##    Suburban 78.02720 44.54184 28.43096
##    Urban    91.97908 52.50628 33.51464
```

---

- The frequency data can also be put directly into a matrix.

```
data <- c(57, 87, 103, 50, 42, 49, 42, 22, 26)
tab <- matrix(data, nrow = 3, ncol = 3)
```

```r
row.names(tab) <- c("Rural", "Suburban", "Urban")
colnames(tab) <- c("Grades", "Popular", "Sports")
tab
```

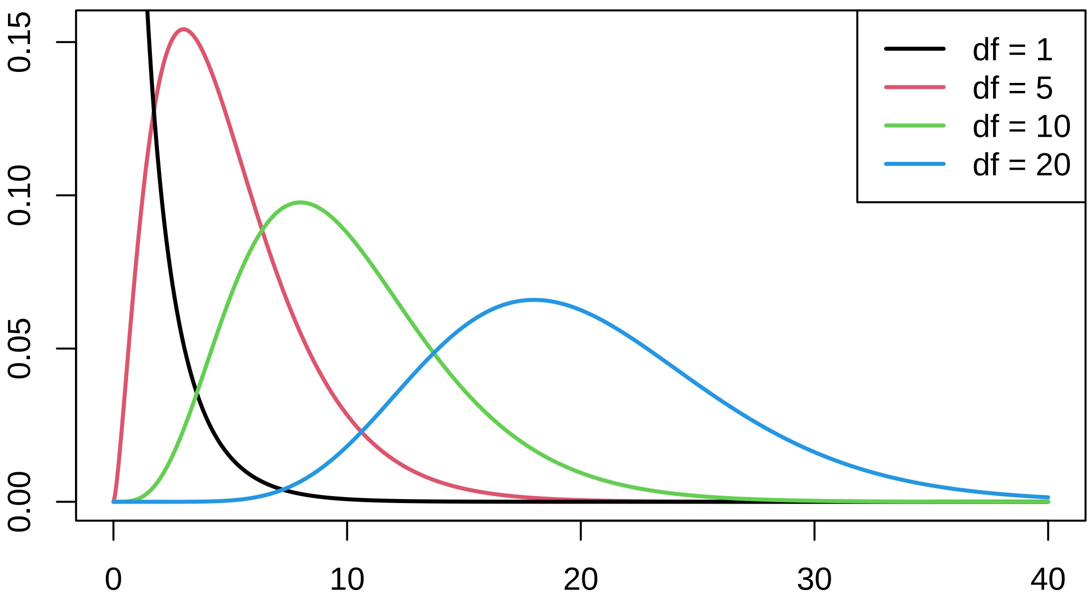
```
##          Grades Popular Sports
## Rural        57      50     42
## Suburban     87      42     22
## Urban       103      49     26
```

```r
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 18.828, df = 4, p-value = 0.0008497
```

### 1.9 The $\chi^2$-distribution

- The $\chi^2$-distribution with $df$ degrees of freedom:
  - Is never negative. And $X^2 = 0$ only happens if $f_e = f_o$.
  - Has mean $\mu = df$
  - Has standard deviation $\sigma = \sqrt{2df}$
  - Is skewed to the right, but approaches a normal distribution when $df$ grows.




### 1.10 Summary

- For the the Chi-squared statistic, $X^2$, to be appropriate we require that the expected values have to be $f_e \geq 5$.
- Now we can summarize the ingredients in the Chi-squared test for independence.

TABLE 8.5: The Five Parts of the Chi-Squared Test of Independence

1. Assumptions: Two categorical variables, random sampling, $f_e \geq 5$ in all cells
2. Hypotheses: $H_0$: Statistical independence of variables
   $H_a$: Statistical dependence of variables
3. Test statistic: $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$, where $f_e = \frac{\text{(Row total)(Column total)}}{\text{Total sample size}}$
4. $P$-value: $P$ = right-tail probability above observed $\chi^2$ value, for chi-squared distribution with $df = (r-1)(c-1)$
5. Conclusion: Report $P$-value
   If decision needed, reject $H_0$ at $\alpha$-level if $P \leq \alpha$

## 1.11 Residual analysis

- If we reject the hypothesis of independence it can be of interest to identify the significant deviations.
- In a given cell in the table, $f_o - f_e$ is the deviation between data and the expected values under the null hypothesis.
- We assume that $f_e \geq 5$.
- If $H_0$ is true, then the standard error of $f_o - f_e$ is given by

$$se = \sqrt{f_e(1 - \textbf{row proportion})(1 - \textbf{column proportion})}$$

- The corresponding $z$-score

$$z = \frac{f_o - f_e}{se}$$

  should in 95% of the cells be between ±2. Values above 3 or below -3 should not appear.
- In `popKids` table cell `Rural and Grade` we got $f_e = 77.0$ and $f_o = 57$. Here **column proportion** $= 0.517$ and **row proportion** $= 149/478 = 0.312$.
- We can then calculate

$$z = \frac{57 - 77}{\sqrt{77(1 - 0.517)(1 - 0.312)}} = -3.95$$

- Compared to the null hypothesis there are way too few rural kids who find grades important.
- In summary: The standardized residuals allow for cell-by-cell ($f_e$ vs $f_o$) comparision.

## 1.12 Residual analysis in R

- In `R` we can extract the standardized residuals from the output of `chisq.test`:

```
tab <- tally(~ Urban.Rural + Goals, data = popKids)
testStat <- chisq.test(tab, correct = FALSE)
testStat$stdres
```

```
##              Goals
## Urban.Rural      Grades     Popular      Sports
##    Rural     -3.9508449   1.3096235   3.5225004
##    Suburban   1.7666608  -0.5484075  -1.6185210
##    Urban      2.0865780  -0.7274327  -1.8186224
```

### 1.13 Cramér's V

- To measure the strength of the association, the Swedish mathematician Harald Cramér developed a measure which is estimated by

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

where r and c are the number of columns and rows in the contingency table and n is the sample size.

- Property:
  - Cramér's V lies between 0(no association) and 1(complete association)
- In the situation with the factors `Goals` and `Urban.Rural` from the dataset `popularKids` we get

$$V = \sqrt{\frac{X^2}{n \cdot \min(r-1, c-1)}} = \sqrt{\frac{18.8}{479 \cdot \min(3-1, 3-1)}} = 0.14,$$

which indicates a weak (but significant) association.

- The function `CramerV` in the package `DescTools` gives you the value and a confidence interval

```
library(DescTools)
```

```
##
## Attaching package: 'DescTools'

## The following object is masked from 'package:mosaic':
##
##     MAD
```

```
CramerV(tab, conf = 0.95, type = "perc")
```

```
##   Cramer V     lwr.ci     upr.ci
## 0.14033592 0.06014641 0.19419139
```

## 2 Ordinal variables

### 2.1 Association between ordinal variables

- For a random sample of black males the General Social Survey in 1996 asked two questions:
  - Q1: What is your yearly income (`income`)?
  - Q2: How satisfied are you with your job (`satisfaction`)?
- Both measurements are on an ordinal scale.

| | VeryD | LittleD | ModerateS | VeryS |
|---|---|---|---|---|
| **$<$ 15k** | 1 | 3 | 10 | 6 |
| **15-25k** | 2 | 3 | 10 | 7 |
| **25-40k** | 1 | 6 | 14 | 12 |
| **$>$ 40k** | 0 | 1 | 9 | 11 |

- We might do a chi-square test to see whether Q1 and Q2 are associated, but the test does not exploit the ordinality.
- We shall consider a test that incorporates ordinality.

## 2.2 Gamma coefficient

- Consider a pair of respondents, where **respondent 1** is below **respondent 2** in relation to Q1.
  - If **respondent 1** is also below **respondent 2** in relation to Q2 then the pair is *concordant*.
  - If **respondent 1** is above **respondent 2** in relation to Q2 then the pair is *disconcordant*.
- Let:

  $C =$ the number of concordant pairs in our sample.

  $D =$ the number of disconcordant pairs in our sample.
- We define the estimated *gamma coefficient*

$$\hat{\gamma} = \frac{C-D}{C+D} = \underbrace{\frac{C}{C+D}}_{concordant\ prop.} - \underbrace{\frac{D}{C+D}}_{discordant\ prop.}$$

## 2.3 Gamma coefficient

- Properties:
  - Gamma lies between -1 og 1
  - The sign tells whether the association is positive or negative
  - Large absolute values correspond to strong association
- The standard error $se(\hat{\gamma})$ on $\hat{\gamma}$ is complicated to calculate, so we leave that to software.
- We can now determine a 95% confidence interval:

$$\hat{\gamma} \pm 1.96 se(\hat{\gamma})$$

  and if zero is contained in the interval, then there is no significant association, when we perform a test with a 5% significance level.

## 2.4 Example

- First, we need to install the package `vcdExtra`, which has the function `GKgamma` for calculating gamma. It also has the dataset on job satisfaction and income built-in:

```
library(vcdExtra)
JobSat
```

```
##           satisfaction
## income    VeryD LittleD ModerateS VeryS
##   < 15k       1       3        10     6
##   15-25k      2       3        10     7
##   25-40k      1       6        14    12
##   > 40k       0       1         9    11
```

```
GKgamma(JobSat, level = 0.90)
```

```
## gamma        : 0.221
## std. error   : 0.117
## CI           : 0.028 0.414
```

- A positive association. Marginally significant at the 10% level, but not so at the 5% level.

# 3 Validation of data

## 3.1 Goodness of fit test

- You have collected a sample and want to know, whether the sample is representative for people living in Hirtshals.
- E.g. whether the distribution of gender, age, or profession in the sample do not differ significantly from the distribution in Hirtshals.
- Actually, you know how to do that for binary variables like gender, but not if you e.g. have 6 agegroups.

## 3.2 Example

- As an example we look at $k$ groups, where data from Hjørring kommune tells us the distribution in Hirtshals is given by the vector
$$\pi = (\pi_1, \dots, \pi_k),$$
where $\pi_i$ is the proportion which belongs to group number $i$, $\quad i = 1, 2 \dots, k$ in Hirtshals.
- Consider the sample represented by the vector:
$$O = (O_1, \dots, O_k),$$
where $O_i$ is the observed number of individuals in group number $i$, $\quad i = 1, 2, \dots, k$.
- The total number of individuals:
$$n = \sum_{i=1}^{k} O_i.$$
- The expected number of individuals in each group, if we have a sample from Hirtshals:
$$E_i = n\pi_i, \;\; i = 1, 2, \dots, k$$

## 3.3 Goodness of fit test

- We will use the following measure to see how far away the observed is from the expected:
$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$
- If this is large we reject the hypothesis that the sample has the same distribution as Hirtshals. The reference distribution is the $\chi^2$ with $k-1$ degrees of freedom.

## 3.4 Example

- Assume we have four groups and that the true distribution is given by:

```
k <- 4
pi_vector <- c(0.3, 0.2, 0.25, 0.25)
```

- Assume that we have the following sample:

```
O_vector <- c(74, 72, 40, 61)
```

- Expected number of individuals in each group:

```
n <- sum(O_vector)
E_vector <- n * pi_vector
E_vector
```

```
## [1] 74.10 49.40 61.75 61.75
```

- $X^2$ statistic:

```r
Xsq = sum((O_vector - E_vector)^2 / E_vector)
Xsq
```

```
## [1] 18.00945
```

- $p$-value:

```r
p_value <- 1 - pchisq(Xsq, df = k-1)
p_value
```

```
## [1] 0.0004378808
```

### 3.5 Test in R

```r
Xsq_test <- chisq.test(O_vector, p = pi_vector)
Xsq_test
```

```
##
##  Chi-squared test for given probabilities
##
## data:  O_vector
## X-squared = 18.009, df = 3, p-value = 0.0004379
```

- As the hypothesis is rejected, we look at the standardized residuals ($z$-scores):

```r
Xsq_test$stdres
```

```
## [1] -0.01388487  3.59500891 -3.19602486 -0.11020775
```

- We conclude that group 1 and 4 is close to true distribution in Hirtshals, but in groups 2 og 3 we have a significant mismatch.