Data collection 1/2

The ASTA team

Contents

1 Data collection

1.1 Motivation

Case

1.2 Data collection

- Getting numbers to report is easy
- Getting sensible and trustworthy numbers to report is orders of magnitude more difficult
- Why important?
  - Difference between meaningless analysis and useful analysis
    * Effect of drugs
    * Economy
    * Sales
    * Climate
    * Energy consumption

1.3 Data collection

Ronald Fisher (1890-1962):

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Said about Fisher:

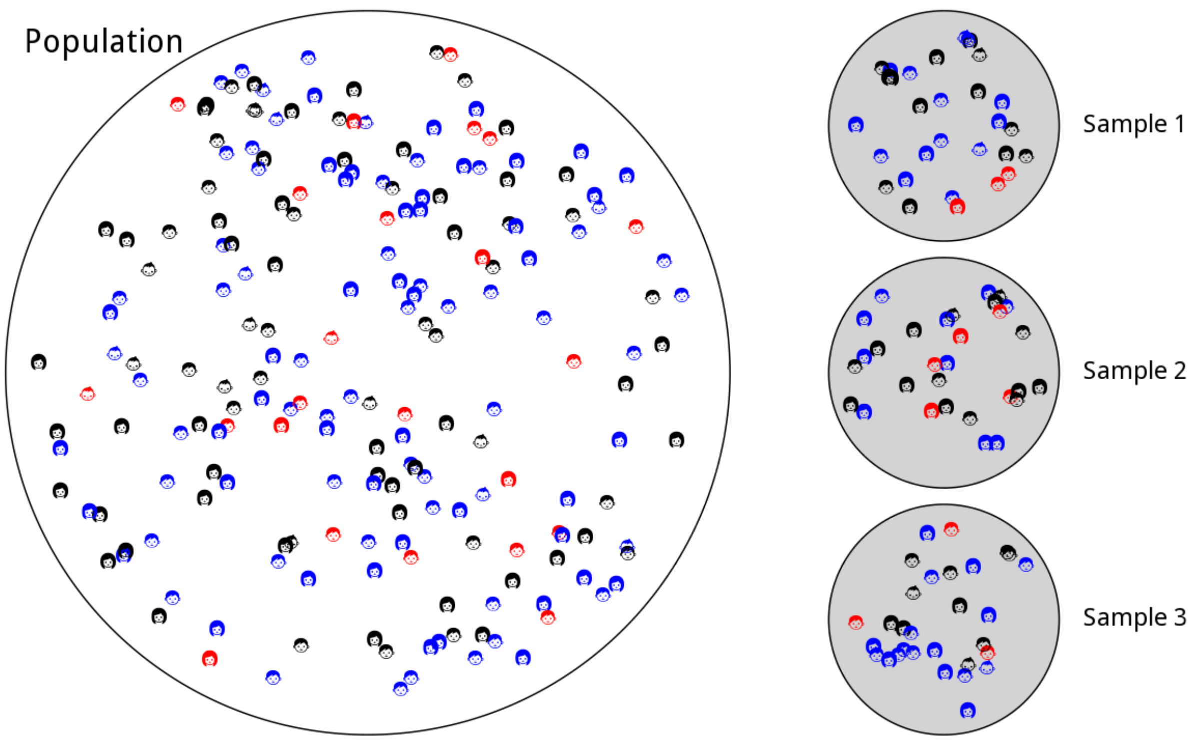
- Anders Hald (1913-2007), Danish statistician: "*a genius who almost single-handedly created the foundations for modern statistical science*"
- Bradley Efron (b. 1938): "*the single most important figure in 20th century statistics*"

1.4 Data collection

- Competences, ideally:
  - Statistics, both conceptually and analyses
  - Data wrangling (loading data; right format for analyses, tables, figures; . . .)
  - Visualizations
  - Knowledge about subject (best with access to experts)
- Not just downloading a spreadsheet!
  - Population vs sample
  - Descriptives of the sample (e.g. mean)
  - Statistical inference about population (how close is sample's mean to population's mean)
- Do collect and analyze data, but know about pitfalls and limitations in generalisability!

2 Population and sample

2.1 Population and sample



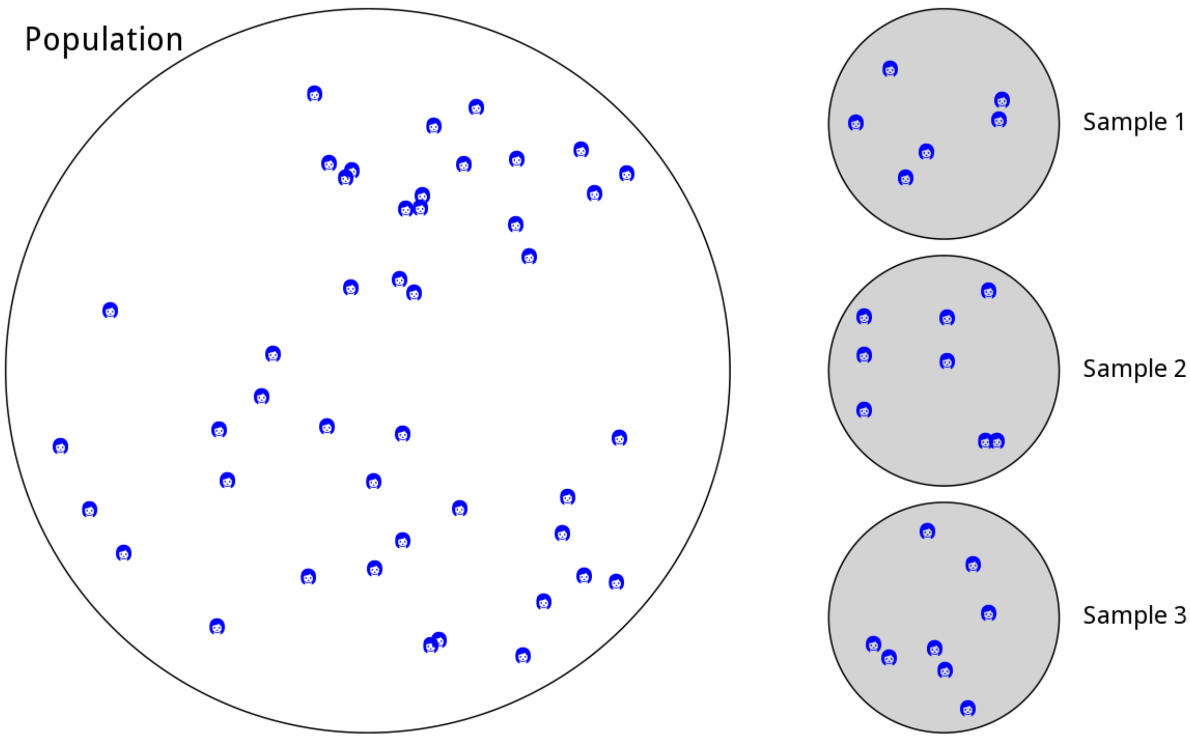
Sample 3 of size $n = 30$:

shape	color	n_sample	p_sample	p_pop	p_diff
baby	black	2	0.07	0.04	-0.02
baby	blue	1	0.03	0.04	0.01
baby	red	0	0.00	0.01	0.01

shape	color	n_sample	p_sample	p_pop	p_diff
man	black	5	0.17	0.12	-0.04
man	blue	8	0.27	0.22	-0.04
man	red	3	0.10	0.08	-0.02
woman	black	3	0.10	0.23	0.13
woman	blue	8	0.27	0.22	-0.05
woman	red	0	0.00	0.02	0.02

- Descriptive vs statistical inference.

2.2 Population and sample



3 Example: United States presidential election, 1936

3.1 Example: United States presidential election, 1936

(Based on Agresti, this and this.)

- Current president: Franklin D. Roosevelt
- Election: Franklin D. Roosevelt vs Alfred Landon (Republican governor of Kansas)
- Literary Digest: magazine with history of accurately predicting winner of past 5 presidential elections

3.2 Example: United States presidential election, 1936

- Literary Digest poll ($\hat{\pi}$ and $1 - \hat{\pi}$): Landon: 57%; Roosevelt: 43%
- Actual results (π and $1 - \pi$): Landon: 38%; Roosevelt: 62%
- Sampling error: 57%-38% = 19%
  - Practically all of the sampling error was the result of **sample bias**
  - Poll size of > 2 mio. individuals participated – extremely large poll

3.3 Example: United States presidential election, 1936

- Mailing list of about 10 mio. names was created
  - Based on every telephone directory, lists of magasine subscribers, rosters of clubs and associations, and other sources
  - Each one of 10 mio. received a mock ballot and asked to return the marked ballot to the magazine
- "respondents who returned their questionnaires represented only that subset of the population with a relatively intense interest in the subject at hand, and as such constitute in no sense a random sample ... it seems clear that the minority of anti-Roosevelt voters felt more strongly about the election than did the pro-Roosevelt majority" (*The American Statistician*, 1976)
- Biases:
  - Selection bias
    - List generated towards middle- and upper-class voters (e.g. 1936 and telephones)
    - Many unemployed (club memberships and magazine subscribers)
  - Non-response bias
    - Only responses from 2.3/2.4 mio out of 10 million people
    - Cannot force people to participate: but mail may be junk (phone, interviews, online, pay/paid, ...)

4 Example: Bullet holes of honor

4.1 Example: Bullet holes of honor

(Based on this.)

- World War II
- Royal Air Force (RAF), UK
  - Lost many planes to German anti-aircraft fire
- Armor up!
  - Where?
  - Count up all the bullet holes in planes that returned from missions
    - Put extra armor in the areas that attracted the most fire

4.2 Example: Bullet holes of honor

- Hungarian-born mathematician Abraham Wald:
  - If a plane makes it back safely with a bunch of bullet holes in its wings: holes in the wings aren't very dangerous

        - **Survivorship bias**
    - Armor up the areas that (on average) don't have any bullet holes
        - They never make it back, apparently dangerous

5 Theory: Biases / sampling

5.1 Biases

Agresti section 2.3:

- Sampling/selection bias
    - Probability sampling: each sample of size n has same probability of being sampled
        - Still problems: undercoverage, groups not represented (inmates, homeless, hospitalized, ...)
    - Non-probability sampling: probability of sample not possible to determine
        - E.g. volunteer sampling
- Response bias
    - E.g. poorly worded, confusing or even order of questions
    - Lying if think socially unacceptable
- Non-response bias
    - Non-response rate high; systematic in non-responses (age, health, believes)

5.2 Sampling

Agresti section 2.4:

- Random sampling schemes:
    - Simple sampling: each possible sample equally probable
    - Systematic sampling
    - Stratified sampling
    - Cluster sampling
    - Multistage sampling
    - ...

6 Theory: Contingency tables

6.1 A contingency table

- We return to the dataset `popularKids`, where we study **association** between 2 **factors**: `Goals` and `Urban.Rural`.
- Based on a sample we make a cross tabulation of the factors and we get a so-called **contingency table** (*krydstabel*).

```
popKids <- read.delim("https://asta.math.aau.dk/datasets?file=PopularKids.txt")
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
```

6.2 A conditional distribution

- Another representation of data is the percent-wise distribution of `Goals` for each level of `Urban.Rural`, i.e. the sum in each row of the table is 100 (up to rounding):

```
tab <- tally(~Urban.Rural + Goals, data = popKids)
addmargins(round(100 * prop.table(tab, 1)),margin = 1:2)
```

```
##             Goals
## Urban.Rural Grades Popular Sports Sum
##    Rural        38      34     28 100
##    Suburban     58      28     15 101
##    Urban        58      28     15 101
##    Sum         154      90     58 302
```

- 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.

7 Independence

7.1 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       500     300    200
##    Suburban    500     300    200
##    Urban       500     300    200
```

- 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:

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

7.2 The Chi-squared test for independence

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

```
n <- margin.table(tab)
pctGoals <- round(100 * margin.table(tab, 2)/n, 1)
pctGoals
```

```
## Goals
##  Grades Popular  Sports
##      52      30      19
```

- 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 (51.7%)  44 (29.5%)  28 (18.8%) 149 (100%)
##    Suburban  78 (51.7%)  44 (29.5%)  28 (18.8%) 151 (100%)
##    Urban     92 (51.7%)  52 (29.5%)  34 (18.8%) 178 (100%)
##    Sum      247 (51.7%) 141 (29.5%)  90 (18.8%) 478 (100%)
```

7.3 Calculation of expected table

```
pctexptab
```

```
##            Goals
## Urban.Rural Grades      Popular     Sports      Sum
##    Rural     77 (51.7%)  44 (29.5%)  28 (18.8%) 149 (100%)
##    Suburban  78 (51.7%)  44 (29.5%)  28 (18.8%) 151 (100%)
##    Urban     92 (51.7%)  52 (29.5%)  34 (18.8%) 178 (100%)
##    Sum      247 (51.7%) 141 (29.5%)  90 (18.8%) 478 (100%)
```

- We note that
  - The relative frequency for a given column is columnTotal divided by tableTotal. For example `Grades`, which is $\frac{247}{478} = 51.7\%$.
  - The expected value in a given cell in the table is then the cell's relative column frequency multiplied by the cell's rowTotal. For example `Rural` and `Grades`: $149 \times 51.7\% = 77.0$.
- This can be summarized to:
  - The expected value in a cell is the product of the cell's rowTotal and columnTotal divided by tableTotal.

7.4 Chi-squared (χ^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      44       28    149
##    Suburban  78      44       28    151
##    Urban     92      52       34    178
##    Sum      247     141       90    478
```

- 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:

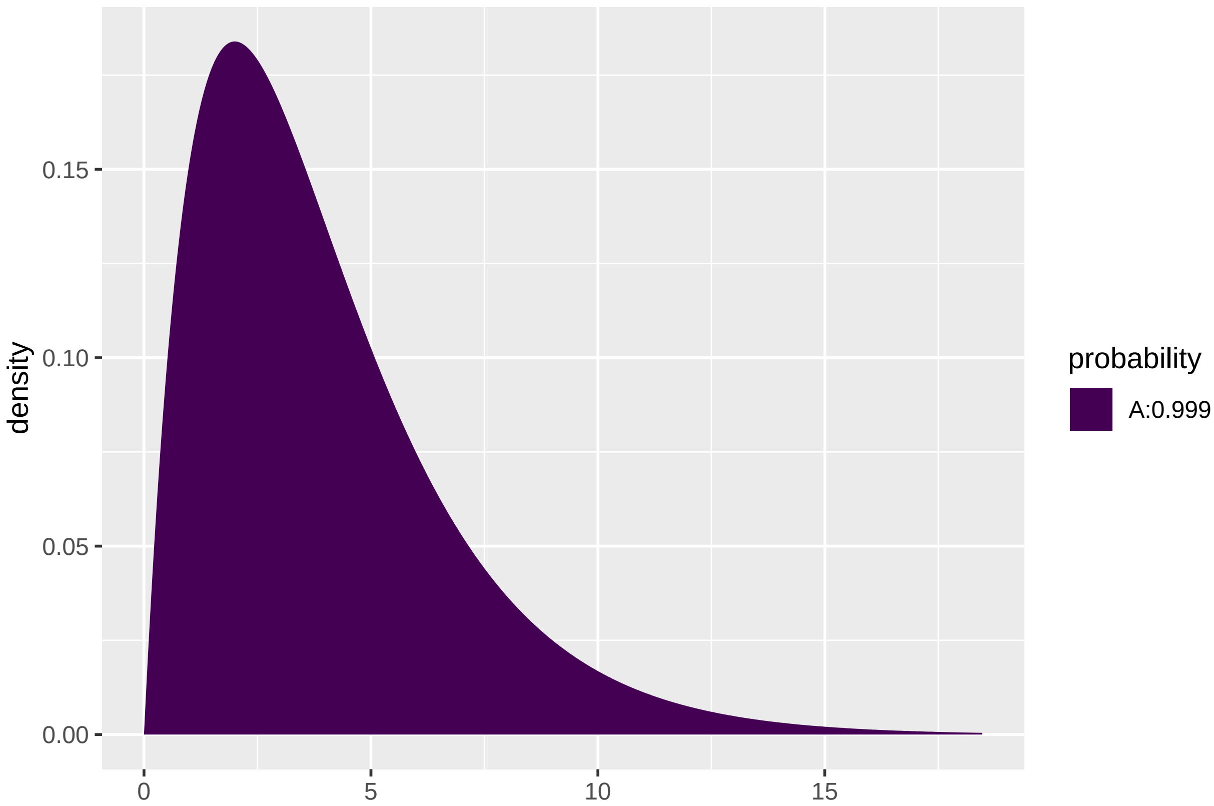
$$X^2_{obs} = \frac{(57-77)^2}{77} + \ldots + \frac{(26-33.5)^2}{33.5} = 18.8$$

- Is this a large distance??

7.5 χ^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 χ^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:

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




```
## [1] 0.00086
```

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

7.6 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 = 20, df = 4, p-value = 8e-04
```

```
testStat$expected
```

```
##             Goals
## Urban.Rural Grades Popular Sports
##    Rural        77      44     28
##    Suburban     78      45     28
##    Urban        92      53     34
```

- 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)
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
```

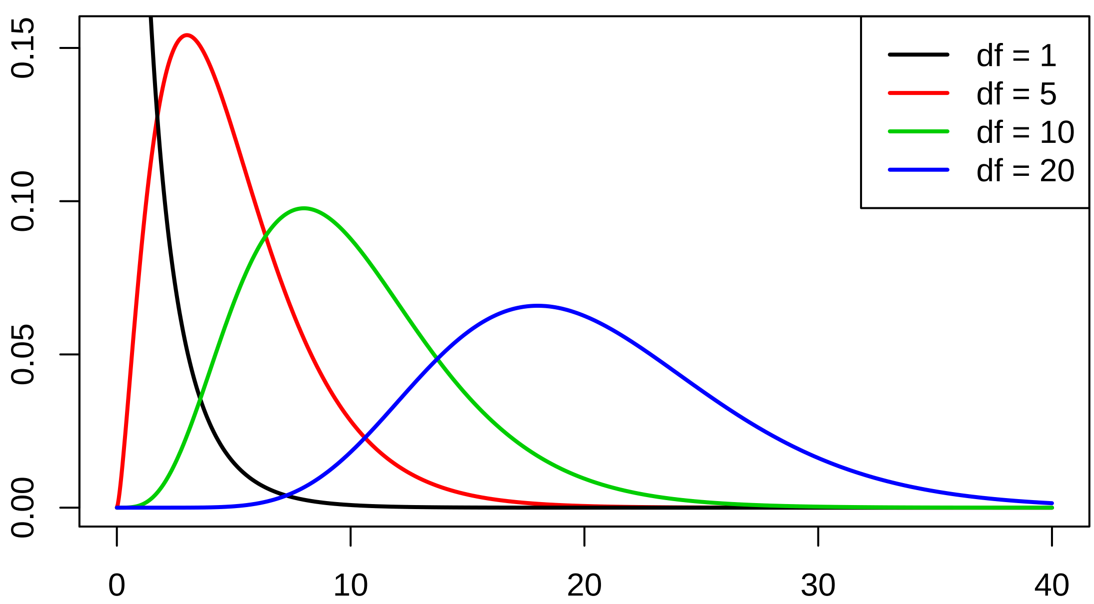
```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 20, df = 4, p-value = 8e-04
```

8 The χ^2-distribution

8.1 The χ^2-distribution

- The χ^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.



9 Agresti - Summary

9.1 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 χ^2 value,
   for chi-squared distribution with $df = (r - 1)(c - 1)$
5. Conclusion: Report P-value
   If decision needed, reject H_0 at α-level if $P \leq \alpha$

10 Standardized residuals

10.1 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

$$se = \sqrt{f_e(1 - \text{rowProportion})(1 - \text{columnProportion})}$$

- 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 columnProportion$= 51.7\%$ and rowProportion$= 149/478 = 31.2\%$.
- 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) comparison.

10.2 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     -3.95    1.31   3.52
##    Suburban   1.77   -0.55  -1.62
##    Urban      2.09   -0.73  -1.82
```

11 Collecting data

11.1 Sources

- Open data
- Questionnaires
  - Google Analyse
  - SurveyXact?
- User panels (often online)
- ...

12 Important take-home messages

12.1 Important take-home messages

- Population vs sample:
  - What is the population?
  - Is the entire population known – is statistics at all needed?
- Sampling
  - Sampling strategy must ensure random sampling
    - Difficult to investigate it afterwards
  - Convenience sampling often used, dangerous!
  - Be honest with yourself, describe problems: Is the sample representative for the target group/population/market segment/...?
- Badly chosen big sample is much worse than a well-chosen small sample
- Watch out for biases
  - Sample/selection bias
  - Response bias
  - Non-response bias
  - (Survivorship bias)
- Data collection
  - Privacy vs necessary information (< 50 or >= 50, age in years, birth date)