Course 02402 Introduction to Statistics Lecture 12:

Inference for proportions

Per Bruun Brockhoff

DTU Compute
Danish Technical University
2800 Lyngby – Denmark
e-mail: perbb@dtu.dk

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Different analysis/data-situations in course 02402

Mean for quantitative data:

- Hypothesis test/CI for one mean (one-sample)
- Hypothesis test/CI for two means (two samples)
- Hypothesis test/CI for several means (K samples)

Different analysis/data-situations in course 02402

Mean for quantitative data:

- Hypothesis test/CI for one mean (one-sample)
- Hypothesis test/CI for two means (two samples)
- Hypothesis test/CI for several means (K samples)

Today: Proportions:

- Hypothesis test/CI for one proportion
- Hypothesis test/CI for two proportions
- Hypothesis test for several proportions
- Hypothesis test for several "multi-categorical" proportions

Estimation of proportions

- Estimation of proportions:

$$\hat{p} = \frac{x}{n}$$

$$\hat{p} \in [0; 1]$$

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Confidence interval for one proportion

Method 7.3

If we have a large sample , then an $(1-\alpha)\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

Confidence interval for one proportion

Method 7.3

If we have a large sample , then an $(1-\alpha)\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

How?

Follows from approximating the binomial distribution by the normal distribution.

Confidence interval for one proportion

Method 7.3

If we have a large sample , then an $(1-\alpha)\%$ confidence interval for p is:

$$\frac{x}{n} - z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}} < p < \frac{x}{n} + z_{1-\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{x}{n}\right)}{n}}$$

How?

Follows from approximating the binomial distribution by the normal distribution.

As a rule of thumb

the normal distribution gives a good approximation of the binomial distrinution if np and $n(1-p)$ are both greater than 15

Confidence interval for one proportion

Mean and variance in binomial distribution, Chapter 2.21

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned}$$

This means that

$$\begin{aligned} E(\hat{p}) = E(\frac{X}{n}) &= \frac{np}{n} = p \\ Var(\hat{p}) = Var(\frac{X}{n}) &= \frac{1}{n^2} Var(X) = \frac{p(1-p)}{n} \end{aligned}$$

Example 1

Left handed:

$$p = \text{proportion of left handed in Denmark}$$

and/or:

Female engineering students:

$$p = \text{Proportion of female engineering students}$$

Example 1

Left handed:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 \Leftrightarrow 0.10 \pm 0.059 \Leftrightarrow [0.041, 0.159]$$

Example 1

Left handed:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{10/100(1-10/100)}{100}} = 0.03$$

$$0.10 \pm 1.96 \cdot 0.03 \Leftrightarrow 0.10 \pm 0.059 \Leftrightarrow [0.041, 0.159]$$

Better "small sample" method - "plus 2-approach":(Remark 7.7)

Use the same formula on $\tilde{x} = 10 + 2 = 12$ and $\tilde{n} = 104$:

$$\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\tilde{n}}} = \sqrt{\frac{12/104(1-12/104)}{104}} = 0.031328$$

$$0.1154 \pm 1.96 \cdot 0.03132 \Leftrightarrow 0.1154 \pm 0.0614 \Leftrightarrow [0.054, 0.177]$$

"Margin of Error" on estimate

Margin of Error

with $(1-\alpha)\%$ confidence becomes:

$$ME = z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

where an estimate of p comes from $p = \frac{x}{n}$

Sample size determination

Method 7.13

If you want a Margin of Error ME with $(1-\alpha)\%$ confidence, then you need the following sample size:

$$n = p(1-p)[\frac{z_{1-\alpha/2}}{ME}]^2$$

Sample size determination

Method 7.13

If you want a Margin of Error ME with $(1-\alpha)\%$ confidence, and you have NO reasonable guess of p, then you need the following sample size:

$$n = \frac{1}{4}\left[\frac{z_{1-\alpha/2}}{ME}\right]^2$$

since the worst case approach is given by: $p = \frac{1}{2}$

Example 1 - continued

Left handed:

Assume that we want $ME = 0.01$ (with $\alpha = 0.05$) - what should n be?

Example 1 - continued

Left handed:

Assume that we want $ME = 0.01$ (with $\alpha = 0.05$) - what should n be?

Assume $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

Example 1 - continued

Left handed:

Assume that we want $ME = 0.01$ (with $\alpha = 0.05$) - what should n be?

Assume $p \approx 0.10$:

$$n = 0.1 \cdot 0.9 \left(\frac{1.96}{0.01} \right)^2 = 3467.4 \approx 3468$$

WITHOUT any assumption on the size of p:

$$n = \frac{1}{4} \left(\frac{1.96}{0.01} \right)^2 = 9604$$

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Steps by hypothesis testing - an overview (Repetition)

1. Formulate the hypotheses and choose the level of significance α (choose the "risk-level")
2. Calculate, using the data, the value of the test statistic
3. Calculate the p-value using the test statistic and the relevant sampling distribution, and compare the p-value and the significance level α and make a conclusion
4. (Alternatively, make a conclusion based on the relevant critical value(s))

Hypothesis test for one proportion

The null and alternative hypothesis for one proportion p:

$$H_0: \quad p = p_0$$
$$H_1: \quad p \neq p_0$$

We either accept H_0 or reject H_0

Calculation of test statistic

Theorem 7.10 and Method 7.11

If the sample size is sufficiently large, we use the test statistic: (If $np_0 > 15$ and $n(1-p_0) > 15$)

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Under the null hypothesis the random variable Z follows a standard normal distribution, $Z \sim N(0, 1^2)$

Finishing the test (Method 7.11)

Find the p-value (evidence against the null hypothesis):

- $2P(Z > |z_{\text{obs}}|)$

Test using the critical value

Alternative hypothesis	reject null hypothesis if
$p \neq p_0$	$z_{\text{obs}} < -z_{1-\alpha/2}$ or $z_{\text{obs}} > z_{1-\alpha/2}$

Example 1 - continued

Is half of all people in Denmark left handed?

$$H_0 : p = 0.5, \;\; H_1 : p \neq 0.5$$

Example 1 - continued

Is half of all people in Denmark left handed?

$$H_0 : p = 0.5, \;\; H_1 : p \neq 0.5$$

Test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1-0.5)}} = -8$$

Example 1 - continued

Is half of all people in Denmark left handed?

$$H_0 : p = 0.5, \;\; H_1 : p \neq 0.5$$

Test statistic:

$$z_{\text{obs}} = \frac{x - np_0}{\sqrt{np_0(1-p_0)}} = \frac{10 - 100 \cdot 0.5}{\sqrt{100 \cdot 0.5(1-0.5)}} = -8$$

p-value:

$$2 \cdot P(Z > 8) = 1.2 \cdot 10^{-15}$$

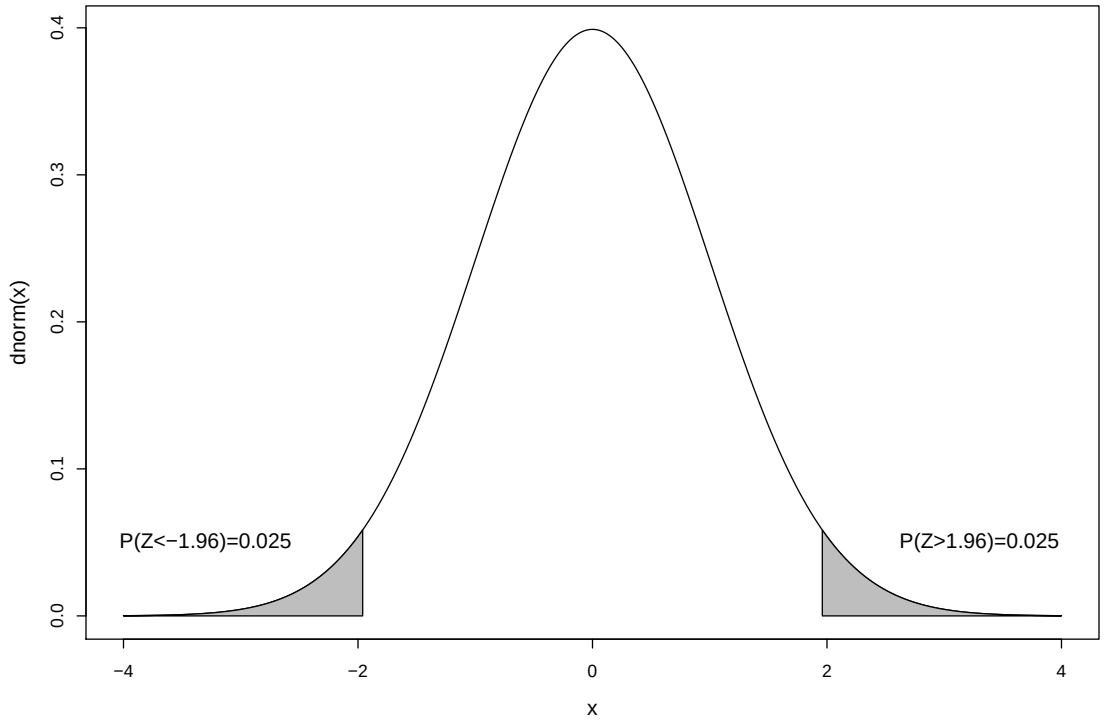
There is very strong evidence against the null hypothesis - we reject this (with $\alpha = 0.05$).

Example 1 - continued

Using the critical value in stead:

$$z_{0.975} = 1.96$$

As $z_{\text{obs}} = -8$ is (much) less than -1.96 we reject the hypothesis.



Agenda

Confidence interval for two proportions

Method 7.15

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \cdot \hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$$

where

$$\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Rule of thumb:

Both $n_i p_i \geq 10$ and $n_i(1-p_i) \geq 10$ for $i = 1, 2$.

Hypothesis test for two proportions, Method 7.18

Two sample proportions hypothesis test

Comparing two proportions (here shown for a two-sided alternative)

$$H_0: \quad p_1 = p_2$$
$$H_1: \quad p_1 \neq p_2$$

The test statistic:

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \quad \textit{where} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

And for large samples:

Use the standard normal distribution again.

Example 2

Is there a relation between ths use of birth control pills and the risk of blood clot in the heart

In a study (USA, 1975) the connection between birth control pills and the risk of blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Is there a relation between ths use of birth control pills and the risk of blood clot in the heart

Carry out a test to check if there is any connection between the use of birth control pills and the risk of blood clot in the heart. Use a significance level of $\alpha = 5\%$

Example 2

In a study (USA, 1975) the connection between birth control pills and the risk of blood clot in the heart was investigated.

	Blood clot	No blood clot
B. C. pill	23	34
No B. C. pill	35	132

Estimates in each sample

$$\hat{p}_1 = \frac{23}{57} = 0.4035, \quad \hat{p}_2 = \frac{35}{167} = 0.2096$$

Common estimate:

$$\hat{p} = \frac{23+35}{57+167} = \frac{58}{224} = 0.2589$$

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Hypothesis test for several proportions

The comparison of c proportions

In some cases we might be interested in determining if two or more binomial distributions have the same parameter p, that is we are interested in testing the null hypothesis:

$$H_0: \quad p_1 = p_2 = \ldots = p_c = p$$

vs. the alternative that the proportions are not equal.

Hypothesis test for several proportions

Table of observed counts for k samples:

	sample 1	sample 2	...	sample c	Total
Succes	x_1	x_2	...	x_c	x
Failure	$n_1 - x_1$	$n_2 - x_2$	...	$n_c - x_c$	$n - x$
Total	n_1	n_2	...	n_c	n

Common (average) estimate:

Under the null hypothesis the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

Hypothesis test for several proportions

Common (average) estimate:

Under the null hypothesis the estimate of p is:

$$\hat{p} = \frac{x}{n}$$

"Use" this common estimate in each group:

If the null hypothesis is true, we expect that the j'th group has e_{1j} successes and e_{2j} failure, where

$$e_{1j} = n_j \cdot \hat{p} = \frac{n_j \cdot x}{n}$$

$$e_{2j} = n_j(1 - \hat{p}) = \frac{n_j \cdot (n - x)}{n}$$

Hypothesis test for several proportions

We fill compute table of EXPECTED counts for k samples:

e_{ij}	sample 1	sample 2	...	sample c	Total
Succes	e_{11}	e_{12}	...	e_{1c}	x
Failure	e_{21}	e_{22}	...	e_{2c}	$n-x$
Total	n_1	n_2	...	n_c	n

General way to find the expected counts in freqency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{(\text{total})}$$

Computation of the test statistic - Method 7.20

The test statistic becomes

$$\chi^2_{\text{obs}} = \sum_{i=1}^{2}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed frequency in cell (i,j) and e_{ij} is the expected frequency in cell (i,j)

Find the p-value or use the critical value - Method 7.20

Sampling distribution for test-statistic:

χ^2-distribution with $(c-1)$ degrees of freedom

Critical value method

If $\chi^2_{\text{obs}} > \chi^2_{\alpha}(c-1)$ the null hypothesis is rejected

Rule of thumb for validity of the test:

All expected values: $e_{ij} \geq 5$.

Example 2 - continued

The OBSERVED values o_{ij}

Observed	Blood clot	No Blood clot
B. C. pill	23	34
No B. C. pill	35	132

Example 2 - continued

Find the EXPECTED values e_{ij}

Expected	Blood clot	No Blood clot	Total
B. C. pill			57
No B. C. pill			167
Total	58	166	224

Example 2 - continued

Use "the rule" for expected vaues four times, e.g.:

$$e_{22} = \frac{167 \cdot 166}{224} = 123.76$$

The EXPECTED values e_{ij}

Expected	Blood clot	No Blood clot	Total
B. C. pill			57
No B. C. pill			167
Total	58	166	224

Example 2 - continued

The test statistic:

$$\chi^2_{\text{obs}} = \frac{(23-14.76)^2}{14.76} + \frac{(34-42.24)^2}{42.24} + \frac{(35-43.24)^2}{43.24} + \frac{(132-123.76)^2}{123.76}$$
$$= 8.33$$

Critical value:

```
qchisq(0.95, 1)
```

```
[1] 3.841
```

Conclusion:

We reject the null hypothesis - there IS a significant higher risk in the BC pill group.

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

Analysis of contingency tables

A 3×3 table - 3 samples, 3-category outcomes

	4 weeks bef	2 weeks bef	1 week bef
Candidate I	79	91	93
Candidate II	84	66	60
Undecided	37	43	47

Are the votes equally distributed?

$$H_0: \quad p_{i1} = p_{i2} = p_{i3}, i = 1, 2, 3.$$

Analysis of contingency tables

A 3×3 table - 1 sample, two 3-category variables:

	bad	average	good
bad	23	60	29
average	28	79	60
good	9	49	63

Is there a dependency between the rows and columns?

$$H_0 : \; p_{ij} = p_{i\cdot} p_{\cdot j}$$

Computation of the test statistic – no matter type of table 7.22

In a contingency table with r rows and c columns, the test statistic is:

$$\chi^2_{\text{obs}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where o_{ij} is the observed value in cell (i,j) and e_{ij} is the expected value in cell (i,j)

General way to find the expected counts in freqency tables:

$$e_{ij} = \frac{(i\text{'th row total}) \cdot (j\text{'th column total})}{(\text{total})}$$

Find p-value or use critical value - Method 7.22

Sampling distribution for test-statistic:

χ^2-distribution with $(r-1)(c-1)$ degrees of freedom

Critical value method

If $\chi^2_{\text{obs}} > \chi^2_\alpha$ with $(r-1)(c-1)$ degrees of freedom the null hypothesis is rejected

Rule of thumb for validity of the test:

All expected values $e_{ij} \geq 5$.

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R

R: prop.test - one proportion

```
# TESTING THE PROBABILITY = 0.5 WITH A TWO-SIDED ALTERNATIVE
# WE HAVE OBSERVED 518 OUT OF 1154
# WITHOUT CONTINUITY CORRECTIONS

prop.test(518, 1154, p = 0.5, correct = FALSE)
```

R: prop.test - two proportions

```
#READING THE TABLE INTO R
pill.study<-matrix(c(23, 34, 35, 132), ncol = 2, byrow = TRUE)
colnames(pill.study) <- c("Blood Clot", "No Clot")
rownames(pill.study) <- c("Pill", "No pill")

# TESTING THAT THE PROBABILITIES FOR THE TWO GROUPS ARE EQUAL
prop.test(pill.study, correct = FALSE)
```

R: chisq.test - two proportions

```
# CHI2 TEST FOR TESTING THE PROBABILITIES FOR THE TWO GROUPS A
chisq.test(pill.study, correct = FALSE)
#IF WE WANT THE EXPECTED NUMBERS SAVE THE TEST IN AN OBJECT
chi <- chisq.test(pill.study, correct = FALSE)
#THE EXPECTED VALUES
chi$expected
```

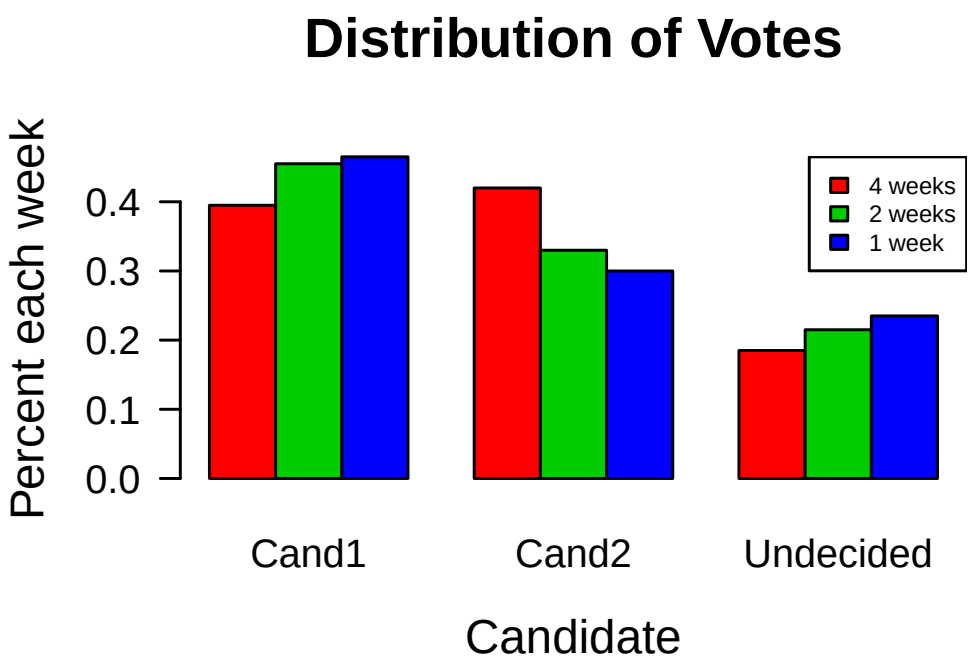
R: chisq.test - contingency tables

```
#READING THE TABLE INTO  R
poll <-matrix(c(79, 91, 93, 84, 66, 60, 37, 43, 47),
              ncol = 3, byrow = TRUE)
colnames(poll) <- c("4 weeks", "2 weeks", "1 week")
rownames(poll) <- c("Cand1", "Cand2", "Undecided")

#COLUMN PERCENTAGES
colpercent<-prop.table(poll, 2)
colpercent
```

R: chisq.test - contingency tables

```
# Plotting percentages
par(mar=c(5,4,4.1,2)+0.1)
barplot(t(colpercent), beside = TRUE, col = 2:4, las = 1,
        ylab = "Percent each week", xlab = "Candidate",
        main = "Distribution of Votes")
legend( legend = colnames(poll), fill = 2:4,"topright", cex = 0.5)
par(mar=c(5,4,4,2)+0.1)
```



R: chisq.test - contingency tables

```
#TESTING SAME DISTRIBUTION IN THE THREE POPULATIONS
chi <- chisq.test(poll, correct = FALSE)
chi

#EXPECTED VALUES
chi$expected
```

Agenda

1. Intro
2. Confidence interval for one proportion
   - Sample size determination (planning)
3. Hypothesis test for one proportion
4. Confidence interval and Hypothesis test for two proportions
5. Hypothesis test for several proportions
6. Analysis of contingency tables
7. R