

# Comparison of two groups

*The ASTA team*

## Contents

0.1	Response variable and explanatory variable . . . . .	1
0.2	Dependent/independent samples . . . . .	1
0.3	Comparison of two means (Independent samples) . . . . .	2
0.4	Comparison of two means (Independent samples) . . . . .	2
0.5	Example: Comparing two means (independent samples) . . . . .	2
0.6	Comparison of two means: confidence interval (independent samples) . . . . .	4
0.7	Comparison of two means: paired $t$ -test (dependent samples) . . . . .	4
<b>1</b>	<b>Comparison of two proportions</b>	<b>6</b>
1.1	Comparison of two proportions . . . . .	6
1.2	Comparison of two proportions: Independent samples . . . . .	6
1.3	Approximate test for comparing two proportions (independent samples) . . . . .	6
1.4	Example: Approximate confidence interval and test for comparing proportions . . . . .	7
1.5	Example: Approximate confidence interval (cont.) . . . . .	7
1.6	Example: $p$ -value (cont.) . . . . .	8
1.7	Automatic calculation in <b>R</b> . . . . .	8
1.8	Fisher's exact test . . . . .	8
1.9	Agresti: Overview of comparison of two groups . . . . .	9

## 0.1 Response variable and explanatory variable

- We conduct an experiment, where we at random choose 50 IT-companies and 50 service companies and measure their profit ratio. Is there association between company type (IT/service) and profit ratio?
- In other words we compare samples from 2 different populations. For each company we register:
  - The binary variable `company type`, which is called **the explanatory variable** and divides data in 2 groups.
  - The quantitative variable `profit ratio`, which is called **the response variable**.

## 0.2 Dependent/independent samples

- In the example with profit ratio of 50 IT-companies and 50 service companies we have **independent samples**, since the same company cannot be in both groups.
- Now, think of another type of experiment, where we at random choose 50 IT-companies and measure their profit ratio in both 2009 and 2010. Then we may be interested in whether there is association between year and profit ratio?
- In this example we have **dependent samples**, since the same company is in both groups.
- Dependent samples may also be referred to as paired samples.

### 0.3 Comparison of two means (Independent samples)

- We consider the situation, where we have two quantitative samples:
  - Population 1 has mean  $\mu_1$ , which is estimated by  $\hat{\mu}_1 = \bar{y}_1$  based on a sample of size  $n_1$ .
  - Population 2 has mean  $\mu_2$ , which is estimated by  $\hat{\mu}_2 = \bar{y}_2$  based on a sample of size  $n_2$ .
  - We are interested in the difference  $\mu_2 - \mu_1$ , which is estimated by  $d = \bar{y}_2 - \bar{y}_1$ .
  - Assume that we can find the **estimated standard error**  $se_d$  of the difference and that this has degrees of freedom  $df$ .
  - Assume that the samples either are large or come from a normal population.
- Then we can construct a
  - confidence interval for the unknown population difference of means  $\mu_2 - \mu_1$  by

$$(\bar{y}_2 - \bar{y}_1) \pm t_{crit} se_d,$$

where the critical  $t$ -score,  $t_{crit}$ , determines the confidence level.

- significance test:
  - \* for the null hypothesis  $H_0 : \mu_2 - \mu_1 = 0$  and alternative hypothesis  $H_a : \mu_2 - \mu_1 \neq 0$ .
  - \* which uses the test statistic:  $t_{obs} = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se_d}$ , that has to be evaluated in a  $t$ -distribution with  $df$  degrees of freedom.

### 0.4 Comparison of two means (Independent samples)

- In the independent samples situation it can be shown that

$$se_d = \sqrt{se_1^2 + se_2^2},$$

where  $se_1$  and  $se_2$  are estimated standard errors for the sample means in populations 1 and 2, respectively.

- We recall, that for these we have  $se = \frac{s}{\sqrt{n}}$ , i.e.

$$se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

where  $s_1$  and  $s_2$  are estimated standard deviations for population 1 and 2, respectively.

- **The degrees of freedom**  $df$  for  $se_d$  can be estimated by a complicated formula, which we will not present here.
- For the confidence interval and the significance test we note that:
  - If both  $n_1$  and  $n_2$  are above 30, then we can use the standard normal distribution ( $z$ -score) rather than the  $t$ -distribution ( $t$ -score).
  - If  $n_1$  or  $n_2$  are below 30, then we let **R** calculate the degrees of freedom and  $p$ -value/confidence interval.

### 0.5 Example: Comparing two means (independent samples)

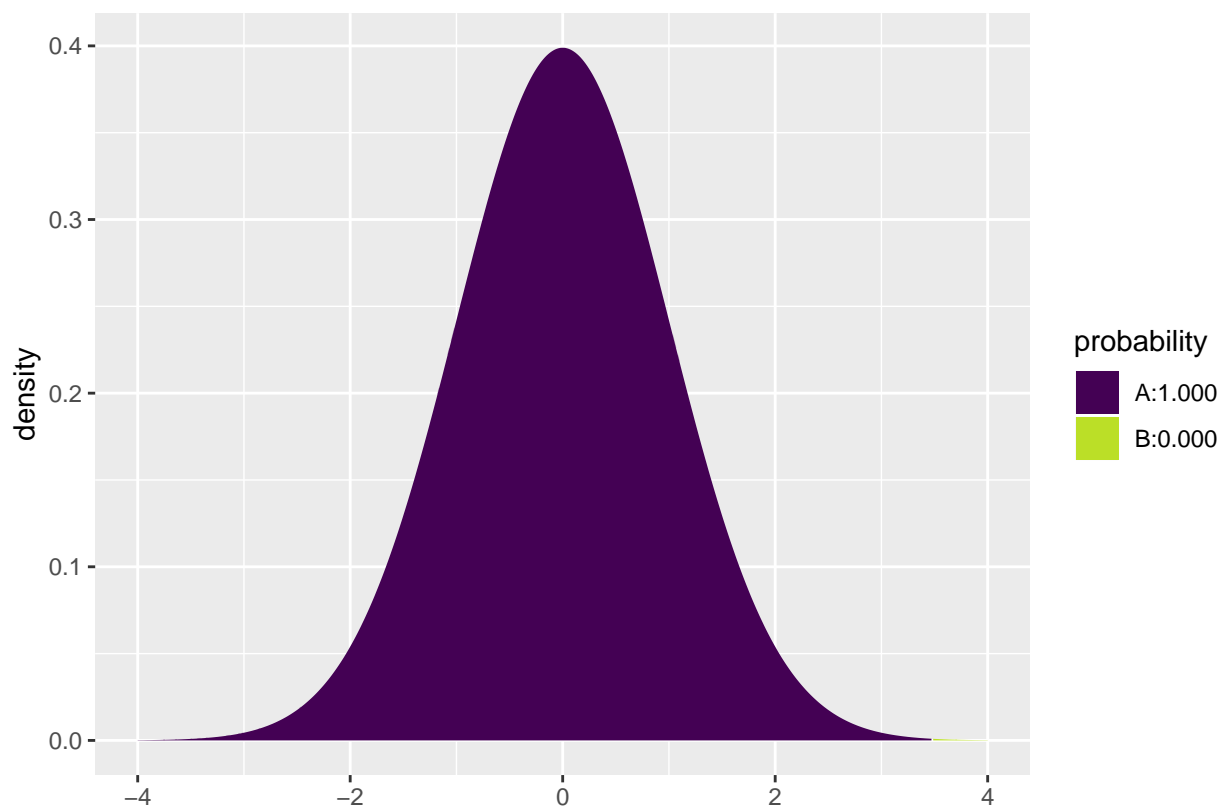
We return to the **Chile** data. We study the association between the variables **sex** and **statusquo** (scale of support for the status-quo). So, we will perform a significance test to test for difference in the mean of **statusquo** for male and females.

```
Chile <- read.delim("https://asta.math.aau.dk/datasets?file=Chile.txt")
library(mosaic)
fv <- favstats(statusquo ~ sex, data = Chile)
fv
```

```
## sex min Q1 median Q3 max mean sd n missing
## 1 F -1.80 -0.975 0.121 1.033 2.02 0.0657 1.003 1368 11
## 2 M -1.74 -1.032 -0.216 0.861 2.05 -0.0684 0.993 1315 6
```

- Difference:  $d = 0.0657 - (-0.0684) = 0.1341$ .
- Estimated standard deviations:  $s_1 = 1.0032$  (females) and  $s_2 = 0.9928$  (males).
- Sample sizes:  $n_1 = 1368$  and  $n_2 = 1315$ .
- Estimated standard error of difference:  $se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{1.0032^2}{1368} + \frac{0.9928^2}{1315}} = 0.0385$ .
- Observed  $t$ -score for  $H_0 : \mu_1 - \mu_2 = 0$  is:  $t_{obs} = \frac{d-0}{se_d} = \frac{0.1341}{0.0385} = 3.4786$ .
- Since both sample sizes are “pretty large” ( $> 30$ ), we can use the  $z$ -score instead of the  $t$ -score for finding the  $p$ -value (i.e. we use the standard normal distribution):

```
1 - pdist("norm", q = 3.4786, xlim = c(-4, 4))
```



```
## [1] 0.0002520202
```

- Then the  $p$ -value is  $2 \cdot 0.00025 = 0.0005$ , so we reject the null hypothesis.
- We can leave all the calculations to **R** by using `t.test`:

```
t.test(statusquo ~ sex, data = Chile)
```

```
##
## Welch Two Sample t-test
##
## data: statusquo by sex
```

```
## t = 3.4786, df = 2678.7, p-value = 0.0005121
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.05849179 0.20962982
## sample estimates:
## mean in group F mean in group M
## 0.06570627 -0.06835453
```

- We recognize the  $t$ -score 3.4786 and the  $p$ -value 0.0005. The estimated degrees of freedom  $df = 2679$  is so large that we can not tell the difference between results obtained using  $z$ -score and  $t$ -score.

## 0.6 Comparison of two means: confidence interval (independent samples)

- We have already found all the ingredients to construct a **confidence interval for  $\mu_2 - \mu_1$** :
  - $d = \bar{y}_2 - \bar{y}_1$  estimates  $\mu_2 - \mu_1$ .
  - $se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$  estimates the standard error of  $d$ .

- Then:

$$d \pm t_{crit} se_d$$

is a confidence interval for  $\mu_2 - \mu_1$ .

- The critical  $t$ -score,  $t_{crit}$  is chosen corresponding to the wanted confidence level. If  $n_1$  and  $n_2$  both are greater than 30, then  $t_{crit} = 2$  yields a confidence level of approximately 95%.

## 0.7 Comparison of two means: paired $t$ -test (dependent samples)

- Experiment:
  - You choose 32 students at random and measure their average reaction time in a driving simulator while they are listening to radio or audio books.
  - Later the same 32 students redo the simulated driving while talking on a cell phone.
- It is interesting to investigate whether or not the fact that you are actively participating in a conversation changes your average reaction time compared to when you are passively listening.
- So we have 2 samples corresponding to with/without phone. In this case we have **dependent** samples, since we have 2 measurement for each student.
- We use the following strategy for analysis:
  - For each student calculate **the change** in average reaction time with and without talking on the phone.
  - The changes  $d_1, d_2, \dots, d_{32}$  are now considered as **ONE** sample from a population with mean  $\mu$ .
  - Test the hypothesis  $H_0 : \mu = 0$  as usual (using a  $t$ -test for testing the mean as in the previous lecture).

### 0.7.1 Reaction time example

- Data is organized in a data frame with 3 variables:
  - **student** (integer – a simple id)
  - **reaction\_time** (numeric – average reaction time in milliseconds)
  - **phone** (factor – yes/no indicating whether speaking on the phone)

```
reaction <- read.delim("https://asta.math.aau.dk/datasets?file=reaction.txt")
head(reaction, n = 3)
```

```
## student reaction_time phone
## 1      1           604    no
## 2      2           556    no
## 3      3           540    no
```

Instead of doing manual calculations we let **R** perform the significance test (using `t.test` with `paired = TRUE` as our samples are paired/dependent):

```
t.test(reaction_time ~ phone, data = reaction, paired = TRUE)
```

```
##
## Paired t-test
##
## data: reaction_time by phone
## t = -5.4563, df = 31, p-value = 5.803e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -69.54814 -31.70186
## sample estimates:
## mean of the differences
## -50.625
```

- With a  $p$ -value of 0.0000058 we reject that speaking on the phone has no influence on the reaction time.
- To understand what is going on, we can manually find the reaction time difference for each student and do a one sample t-test on this difference:

```
yes <- subset(reaction, phone == "yes")
no <- subset(reaction, phone == "no")
reaction_diff <- data.frame(student = no$student, yes = yes$reaction_time, no = no$reaction_time)
reaction_diff$diff <- reaction_diff$yes - reaction_diff$no
head(reaction_diff)
```

```
## student yes no diff
## 1      1 636 604 32
## 2      2 623 556 67
## 3      3 615 540 75
## 4      4 672 522 150
## 5      5 601 459 142
## 6      6 600 544 56
```

```
t.test(~ diff, data = reaction_diff)
```

```
##
## One Sample t-test
##
## data: diff
## t = 5.4563, df = 31, p-value = 5.803e-06
```

```

## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 31.70186 69.54814
## sample estimates:
## mean of x
## 50.625

```

## 1 Comparison of two proportions

### 1.1 Comparison of two proportions

- We consider the situation, where we have two qualitative samples and we investigate whether a given property is present or not:
  - Let the proportion of population 1 which has the property be  $\pi_1$ , which is estimated by  $\hat{\pi}_1$  based on a sample of size  $n_1$ .
  - Let the proportion of population 2 which has the property be  $\pi_2$ , which is estimated by  $\hat{\pi}_2$  based on a sample of size  $n_2$ .
  - We are interested in the difference  $\pi_2 - \pi_1$ , which is estimated by  $d = \hat{\pi}_2 - \hat{\pi}_1$ .
  - Assume that we can find the **estimated standard error**  $se_d$  of the difference.
- Then we can construct
  - an approximate confidence interval for the difference,  $\pi_2 - \pi_1$ .
  - a significance test.

### 1.2 Comparison of two proportions: Independent samples

- In the situation where we have independent samples we know that

$$se_d = \sqrt{se_1^2 + se_2^2},$$

where  $se_1$  and  $se_2$  are the estimated standard errors for the sample proportion in population 1 and 2, respectively.

- We recall, that these are given by  $se = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$ , i.e.

$$se_d = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}.$$

- A (approximate) confidence interval for  $\pi_2 - \pi_1$  is obtained by the usual construction:

$$(\hat{\pi}_2 - \hat{\pi}_1) \pm z_{crit} se_d,$$

where the critical  $z$ -score determines the confidence level.

### 1.3 Approximate test for comparing two proportions (independent samples)

- We consider the null hypothesis  $H_0: \pi_1 = \pi_2$  (equivalently  $H_0: \pi_1 - \pi_2 = 0$ ) and the alternative hypothesis  $H_a: \pi_1 \neq \pi_2$ .

- Assuming  $H_0$  is true, we have a common proportion  $\pi$ , which is estimated by

$$\hat{\pi} = \frac{n_1 \hat{\pi}_1 + n_2 \hat{\pi}_2}{n_1 + n_2},$$

i.e. we aggregate the populations and calculate the relative frequency of the property (with other words: we estimate the proportion,  $\pi$ , as if the two samples were one).

- Rather than using the estimated standard error of the difference from previous, we use the following that holds under  $H_0$ :

$$se_0 = \sqrt{\hat{\pi}(1 - \hat{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

- The observed test statistic/ $z$ -score for  $H_0$  is then:

$$z_{obs} = \frac{(\hat{\pi}_2 - \hat{\pi}_1) - 0}{se_0},$$

which is evaluated in the standard normal distribution.

- The  $p$ -value is calculated in the usual way.

**WARNING:** The approximation is only good, when  $n_1 \hat{\pi}$ ,  $n_1(1 - \hat{\pi})$ ,  $n_2 \hat{\pi}$ ,  $n_2(1 - \hat{\pi})$  all are greater than 5.

## 1.4 Example: Approximate confidence interval and test for comparing proportions

We return to the `Chile` dataset. We make a new binary variable indicating whether the person intends to vote no or something else (and we remember to tell `R` that it should think of this as a grouping variable, i.e. a `factor`):

```
Chile$voteNo <- relevel(factor(Chile$vote == "N"), ref = "TRUE")
```

We study the association between the variables `sex` and `voteNo`:

```
tab <- tally(~ sex + voteNo, data = Chile, useNA = "no")
tab
```

```
##   voteNo
## sex TRUE FALSE
##  F  363   946
##  M  526   697
```

This gives us all the ingredients needed in the hypothesis test:

- Estimated proportion of men that vote no:  $\hat{\pi}_1 = \frac{526}{526+697} = 0.430$
- Estimated proportion of women that vote no:  $\hat{\pi}_2 = \frac{363}{363+946} = 0.277$

## 1.5 Example: Approximate confidence interval (cont.)

- Estimated difference:

$$d = \hat{\pi}_2 - \hat{\pi}_1 = 0.277 - 0.430 = -0.153$$

- Standard error of difference:

$$se_d = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

$$= \sqrt{\frac{0.430(1 - 0.430)}{1223} + \frac{0.277(1 - 0.277)}{1309}} = 0.0188.$$

- Approximate 95% confidence interval for difference:

$$d \pm 1.96 \cdot se_d = (-0.190, -0.116).$$

## 1.6 Example: *p*-value (cont.)

- Estimated common proportion:

$$\hat{\pi} = \frac{1223 \times 0.430 + 1309 \times 0.277}{1309 + 1223} = \frac{526 + 363}{1309 + 1223} = 0.351.$$

- Standard error of difference when  $H_0 : \pi_1 = \pi_2$  is true:

$$se_0 = \sqrt{\hat{\pi}(1 - \hat{\pi}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)} = 0.0190.$$

- The observed test statistic/*z*-score:

$$z_{obs} = \frac{d}{se_0} = -8.06.$$

- The test for  $H_0$  against  $H_a : \pi_1 \neq \pi_2$  yields a *p*-value that is practically zero, i.e. we can reject that the proportions are equal.

## 1.7 Automatic calculation in R

```
Chile2 <- subset(Chile, !is.na(voteNo))
prop.test(voteNo ~ sex, data = Chile2, correct = FALSE)
```

```
##
## 2-sample test for equality of proportions without continuity
## correction
##
## data:  tally(voteNo ~ sex)
## X-squared = 64.777, df = 1, p-value = 8.389e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.1896305 -0.1159275
## sample estimates:
##  prop 1    prop 2
## 0.2773109 0.4300899
```

## 1.8 Fisher's exact test

- If  $n_1\hat{\pi}$ ,  $n_1(1 - \hat{\pi})$ ,  $n_2\hat{\pi}$ ,  $n_2(1 - \hat{\pi})$  are not all greater than 5, then the approximate test cannot be trusted. Instead you can use Fisher's exact test:



```
fisher.test(tab)
```

```
##
## Fisher's Exact Test for Count Data
##
## data: tab
## p-value = 1.04e-15
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 0.4292768 0.6021525
## sample estimates:
## odds ratio
## 0.5085996
```

- Again the  $p$ -value is seen to be extremely small, so we definitely reject the null hypothesis of equal vote proportions for women and men.

## 1.9 Agresti: Overview of comparison of two groups

**TABLE 7.10:** Summary of Comparison Methods for Two Groups, for Independent Random Samples

	Type of Response Variable	
	Categorical	Quantitative
<b>Estimation</b>		
1. Parameter	$\pi_2 - \pi_1$	$\mu_2 - \mu_1$
2. Point estimate	$\hat{\pi}_2 - \hat{\pi}_1$	$\bar{y}_2 - \bar{y}_1$
3. Standard error	$se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$	$se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
4. Confidence interval	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$	$(\bar{y}_2 - \bar{y}_1) \pm t(se)$
<b>Significance testing</b>		
1. Assumptions	Randomization $\geq 10$ observations in each category, for each group	Randomization Normal population dist.'s (robust, especially for large $n$ 's)
2. Hypotheses	$H_0: \pi_1 = \pi_2$ $(\pi_2 - \pi_1 = 0)$ $H_a: \pi_1 \neq \pi_2$	$H_0: \mu_1 = \mu_2$ $(\mu_2 - \mu_1 = 0)$ $H_a: \mu_1 \neq \mu_2$
3. Test statistic	$z = \frac{\hat{\pi}_2 - \hat{\pi}_1}{se_0}$	$t = \frac{\bar{y}_2 - \bar{y}_1}{se}$
4. $P$ -value	Two-tail probability from standard normal or $t$ (Use one tail for one-sided alternative)	