ASTA

The ASTA team

Contents

1 Statistical inference: Hypothesis and test

1.1 Concept of hypothesis

- A **hypothesis** is a statement about a given population. Usually it is stated as a population parameter having a given value or being in a certain interval.
- Examples:
  - Quality control of products: The hypothesis is that the products e.g. have a certain weight, a given power consumption or a minimal durability.
  - Scientific hypothesis: There is no dependence between a company's age and level of return.

1.2 Significance test

- A significance test is used to investigate, whether data is contradicting the hypothesis or not.
- If the hypothesis says that a parameter has a certain value, then the test should tell whether the sample estimate is "far" away from this value.
- For example:
  - Waiting times in a queue. We sample n customers and count how many that have been waiting more than 5 minutes. The company policy is that at most 10% of the customers should wait more than 5 minutes. In a sample of size $n = 32$ we observe 4 with waiting time above 5 minutes, i.e. the estimated proportion is $\hat{\pi} = \frac{4}{32} = 12.5\%$. Is this "much more" than (i.e. significantly different from) 10%?
  - The blood alcohol level of a student is measured 4 times with the values $0.504, 0.500, 0.512, 0.524$, i.e. the estimated mean value is $\bar{y} = 0.51$. Is this "much different" than a limit of 0.5?

1.3 Null and alternative hypothesis

- **The null hypothesis** - denoted H_0 - usually specifies that a population parameter has some given value. E.g. if μ is the mean blood alcohol level we can state the null hypothesis
  - $H_0 : \mu = 0.5$.
- The **alternative hypothesis** - denoted H_a - usually specifies that the population parameter is contained in a given set of values different than the null hypothesis. E.g. if μ again is the population mean of a blood alcohol level measurement, then
  - the null hypothesis is $H_0 : \mu = 0.5$
  - the alternative hypothesis is $H_a : \mu \neq 0.5$.

1.4 Test statistic

- We consider a population parameter μ and write the null hypothesis

$$H_0 : \mu = \mu_0,$$

  where μ_0 is a known number, e.g. $\mu_0 = 0.5$.
- Based on a sample we have an estimate $\hat{\mu}$.
- A **test statistic** T will typically depend on $\hat{\mu}$ and μ_0 (we may write this as $T(\hat{\mu}, \mu_0)$) and measures "how far from μ_0 is $\hat{\mu}$?"
- Often we use $T(\hat{\mu}, \mu_0)$ = "the number of standard deviations from $\hat{\mu}$ to μ_0".
- For example it would be very unlikely to be more than 3 standard deviations from μ_0, i.e. in that case μ_0 is probably not the correct value of the population parameter.

1.5 P-value

- We consider
  - H_0: a null hypothesis.
  - H_a: an alternative hypothesis.
  - T: a test statistic, where the value calculated based on the current sample is denoted t_{obs}.
- To investigate the plausibility of H_0, we measure the evidence against H_0 by the so-called p-value:
  - The p-value is the probability of observing a more extreme value of T (if we were to repeat the experiment) than t_{obs} *under the assumption that* H_0 *is true.*
  - "Extremity" is measured relative to the alternative hypothesis; a value is considered extreme if it is "far from" H_0 and "closer to" H_a.
  - If the p-value is small then there is a small probability of observing t_{obs} if H_0 is true, and thus H_0 is not very probable for our sample and we put more support in H_a, so:

The smaller the p-value, the less we trust H_0.

- What is a small p-value? If it is below 5% we say it is **significant** at the 5% level.

1.6 Significance level

- We consider
  - H_0: a null hypothesis.
  - H_a: an alternative hypothesis.
  - T: a test statistic, where the value calculated based on the current sample is denoted t_{obs} and the corresponding p-value is p_{obs}.
- Small values of p_{obs} are critical for H_0.
- In practice it can be necessary to decide whether or not we are going to reject H_0.
- The decision can be made if we previously have decided on a so-called α**-level**, where
  - α is a given percentage
  - we reject H_0, if p_{obs} is less than or equal to α
  - α is called the **significance level** of the test
  - typical choices of α are 5% or 1%.

1.7 Significance test for mean

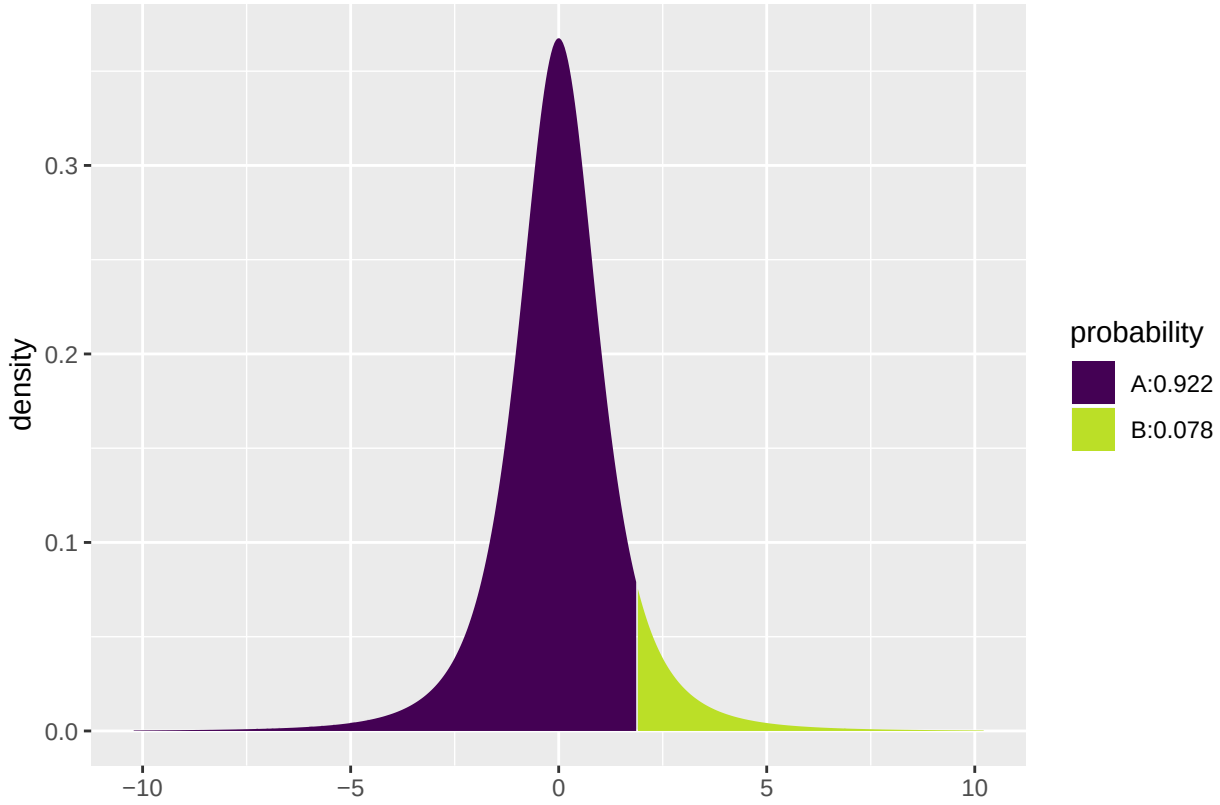
1.7.1 Two-sided t-test for mean:

- We assume that data is a sample from $\texttt{norm}(\mu, \sigma)$.
- The estimates of the population parameters are $\hat{\mu} = \bar{y}$ and $\hat{\sigma} = s$ based on n observations.
- Null hypothesis: $H_0: \ \mu = \mu_0$, where μ_0 is a known value.
- **Two-sided alternative hypothesis**: $H_a: \ \mu \neq \mu_0$.
- Observed test statistic: $t_{obs} = \frac{\bar{y} - \mu_0}{se}$, where $se = \frac{s}{\sqrt{n}}$.
- I.e. t_{obs} measures, how many standard deviations (with $\pm$ sign) the empirical mean lies away from μ_0.
- If H_0 is true, then t_{obs} is an observation from the t-distribution with $df = n - 1$.
- P-value: Values bigger than $|t_{obs}|$ or less than $-|t_{obs}|$ puts more support in H_a than H_0.
- The p-value = 2 x "upper tail probability of $|t_{obs}|$". The probability is calculated in the t-distribution with df degrees of freedom.

1.7.2 Example: Two-sided t-test

- Blood alcohol level measurements: $0.504, 0.500, 0.512, 0.524$.
- These are assumed to be a sample from a normal distribution.
- We calculate
  - $\bar{y} = 0.51$ and $s = 0.0106$
  - $se = \frac{s}{\sqrt{n}} = \frac{0.0106}{\sqrt{4}} = 0.0053$.
  - $H_0 : \mu = 0.5$, i.e. $\mu_0 = 0.5$.
  - $t_{obs} = \frac{\bar{y} - \mu_0}{se} = \frac{0.51 - 0.5}{0.0053} = 1.89$.
- So we are almost 2 standard deviations from 0.5. Is this extreme in a t-distribution with 3 degrees of freedom?

```
library(mosaic)
1 - pdist("t", q = 1.89, df = 3)
```




```
## [1] 0.07757725
```

- The p-value is $2 \cdot 0.078$, i.e. more than 15%. On the basis of this we do not reject H_0.

1.8 One-sided t-test for mean

The book also discusses one-sided t-tests for the mean, but we will not use those in the course.

1.9 Agresti: Overview of t-test

TABLE 6.3: The Five Parts of Significance Tests for Population Means

1. **Assumptions**
   Quantitative variable
   Randomization
   Normal population (robust, especially for two-sided H_a, large n)
2. **Hypotheses**
   H_0: $\mu = \mu_0$
   H_a: $\mu \neq \mu_0$ (or H_a: $\mu > \mu_0$ or H_a: $\mu < \mu_0$)
3. **Test statistic**
   $t = \dfrac{\bar{y} - \mu_0}{se}$ where $se = \dfrac{s}{\sqrt{n}}$
4. ***P*-value**
   In t curve, use
   P = Two-tail probability for H_a: $\mu \neq \mu_0$
   P = Probability to right of observed t-value for H_a: $\mu > \mu_0$
   P = Probability to left of observed t-value for H_a: $\mu < \mu_0$
5. **Conclusion**
   Report P-value. Smaller P provides stronger evidence against H_0 and supporting H_a. Can reject H_0 if $P \leq \alpha$-level.

1.10 Significance test for proportion

- Consider a sample of size n, where we observe whether a given property is present or not.
- The relative frequency of the property in the population is π, which is estimated by $\hat{\pi}$.
- Null hypothesis: $H_0 : \pi = \pi_0$, where π_0 is a known number.
- **Two-sided alternative** hypothesis: $H_a : \pi \neq \pi_0$.
- *If H_0 is true* the standard error for $\hat{\pi}$ is given by $se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}}$.
- Observed test statistic: $z_{obs} = \frac{\hat{\pi}-\pi_0}{se_0}$
- I.e. z_{obs} measures, how many standard deviations (with $\pm$ sign) there is from $\hat{\pi}$ to π_0.

1.10.1 Approximate test

- If both $n\hat{\pi}$ and $n(1-\hat{\pi})$ are larger than 15 we know from previously that $\hat{\pi}$ follows a normal distribution (approximately), i.e.
  - If H_0 is true, then z_{obs} is an observation from the standard normal distribution.
- P-value for **two-sided** test: Values greater than $|z_{obs}|$ or less than $-|z_{obs}|$ point more towards H_a than H_0.
- The p-value=2 x "upper tail probability for $|z_{obs}|$". The probability is calculated in the standard normal distribution.

1.10.2 Example: Approximate test

- We consider a study from Florida Poll 2006:
  - In connection with problems financing public service a random sample of 1200 individuals were asked whether they preferred less service or tax increases.
  - 52% preferred tax increases. Is this enough to say that the proportion is significantly different from fifty-fifty?
- Sample with $n = 1200$ observations and estimated proportion $\hat{\pi} = 0.52$.
- Null hypothesis $H_0 : \pi = 0.5$.
- Alternative hypothesis $H_a : \pi \neq 0.5$.
- Standard error $se_0 = \sqrt{\frac{\pi_0(1-\pi_0)}{n}} = \sqrt{\frac{0.5 \times 0.5}{1200}} = 0.0144$
- Observed test statistic $z_{obs} = \frac{\hat{\pi}-\pi_0}{se_0} = \frac{0.52-0.5}{0.0144} = 1.39$
- "upper tail probability for 1.39" in the standard normal distribution is 0.0823, i.e. we have a p-value of $2 \cdot 0.0823 \approx 16\%$.
- Conclusion: There is not sufficient evidence to reject H_0, i.e. we do not reject that the preference in the population is fifty-fifty.
- Note, the above calculations can also be performed automatically in **R** by (a little different results due to rounding errors in the manual calculation):

```
count <- 1200 * 0.52 # number of individuals preferring tax increase
prop.test(x = count, n = 1200, correct = F)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  count out of 1200
## X-squared = 1.92, df = 1, p-value = 0.1659
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4917142 0.5481581
## sample estimates:
##    p
## 0.52
```

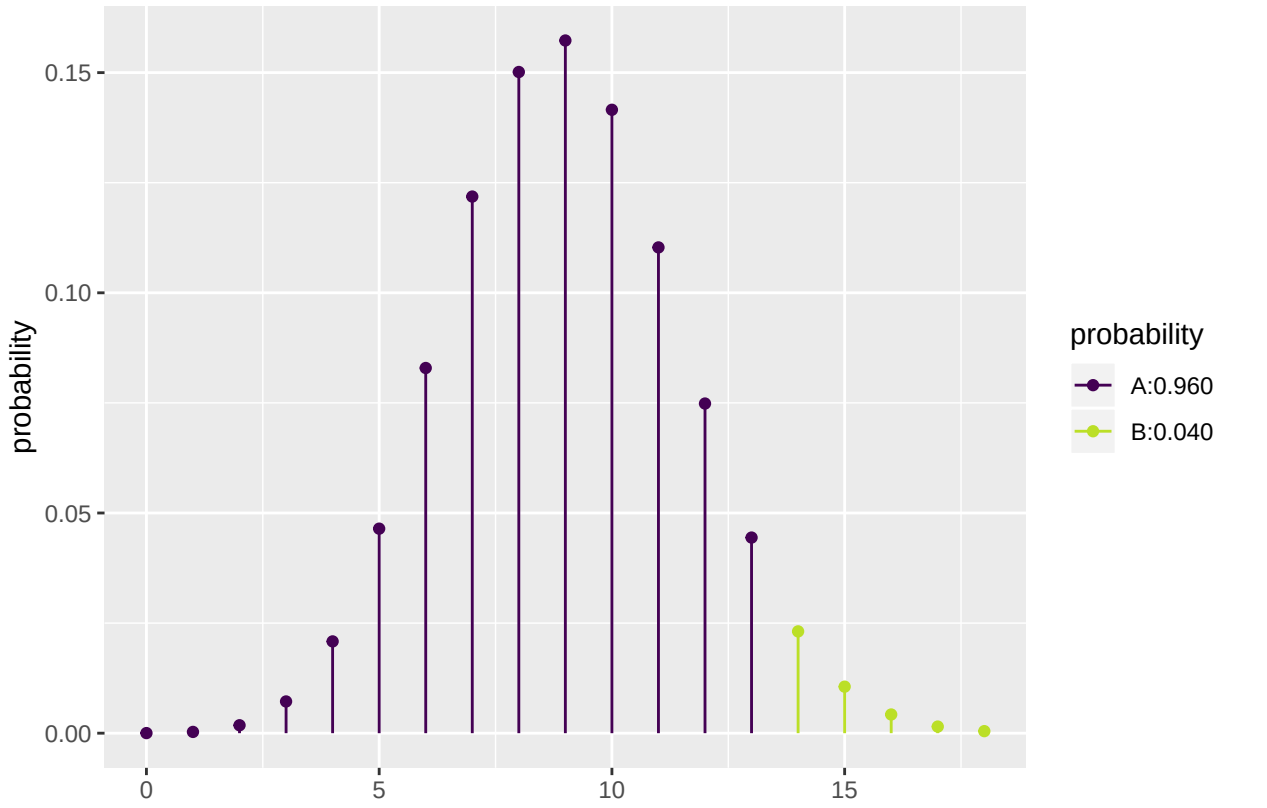
1.10.3 Binomial (exact) test

- Consider again a sample of size n, where we observe whether a given property is present or not.
- The relative frequency of the property in the population is π, which is estimated by $\hat{\pi}$.
- Let $y_+ = n\hat{\pi}$ be the frequency (total count) of the property in the sample.
- It can be shown that y_+ follows the **binomial distribution** with size parameter n and success probability π. We use $Bin(n, \pi)$ to denote this distribution.
- Null hypothesis: $H_0 : \pi = \pi_0$, where π_0 is a known number.
- Alternative hypothesis: $H_a : \pi \neq \pi_0$, where π_0 is a known number.
- P-value for **two-sided** binomial test:
  - If $y_+ \geq n\pi_0$: 2 x "upper tail probability for y_+" in the $Bin(n, \pi_0)$ distribution.
  - If $y_+ < n\pi_0$: 2 x "lower tail probability for y_+" in the $Bin(n, \pi_0)$ distribution.

1.10.4 Example: Binomial test

- Experiment with $n = 30$, where we have $y_+ = 14$ successes.
- We want to test $H_0 : \pi = 0.3$ vs. $H_a : \pi \neq 0.3$.
- Since $y_+ > n\pi_0 = 9$ we use the upper tail probability corresponding to the sum of the height of the red lines to the right of 14 in the graph below. (Notice, the graph continues on the right hand side to $n = 30$, but it has been cut off for illustrative purposes.)
- The upper tail probability from 14 and up (i.e. greater than 13) is:

```
lower_tail <- pdist("binom", q = 13, size = 30, prob = 0.3)
```




```
1 - lower_tail
```

```
## [1] 0.04005255
```

- The two-sided p-value is then 2 x 0.04 = 0.08.

1.10.5 Binomial test in R

- We return to the Chile data, where we again look at the variable `sex`.
- Let us test whether the proportion of females is different from 50 %, i.e., we look at $H_0 : \pi = 0.5$ and $H_a : \pi \neq 0.5$, where π is the unknown population proportion of females.

```
Chile <- read.delim("https://asta.math.aau.dk/datasets?file=Chile.txt")
binom.test( ~ sex, data = Chile, p = 0.5, conf.level = 0.95)
```

```
##
##
##
## data:  Chile$sex  [with success = F]
## number of successes = 1379, number of trials = 2700, p-value =
## 0.2727
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4916971 0.5297610
## sample estimates:
## probability of success
##              0.5107407
```

- The p-value for the binomial exact test is 27%, so there is no significant difference between the proportion of males and females.
- The approximate test has a p-value of 26%, which can be calculated by the command

```
prop.test( ~ sex, data = Chile, p = 0.5, conf.level = 0.95, correct = FALSE)
```

(note the additional argument `correct = FALSE`).

1.11 Agresti: Overview of tests for mean and proportion

TABLE 6.7: Summary of Significance Tests for Means and Proportions

Parameter	Mean	Proportion
1. Assumptions	Random sample, quantitative variable normal population	Random sample, categorical variable null expected counts at least 10
2. Hypotheses	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$ $H_a: \mu > \mu_0$ $H_a: \mu < \mu_0$	$H_0: \pi = \pi_0$ $H_a: \pi \neq \pi_0$ $H_a: \pi > \pi_0$ $H_a: \pi < \pi_0$
3. Test statistic	$t = \frac{\bar{y} - \mu_0}{se}$ with $se = \frac{s}{\sqrt{n}}, df = n - 1$	$z = \frac{\hat{\pi} - \pi_0}{se_0}$ with $se_0 = \sqrt{\pi_0(1 - \pi_0)/n}$
4. *P*-value	Two-tail probability in sampling distribution for two-sided test ($H_0: \mu \neq \mu_0$ or $H_a: \pi \neq \pi_0$); One-tail probability for one-sided test	
5. Conclusion	Reject H_0 if *P*-value $\leq \alpha$-level such as 0.05	

1.12 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**.

1.13 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.

1.14 Comparison of two means (Independent samples)

- We consider the situation, where we have two quantitative samples:
  - Population 1 has mean μ_1, which is estimated by $\hat{\mu}_1 = \bar{y}_1$ based on a sample of size n_1.
  - Population 2 has mean μ_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.

1.15 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.

1.16 Example: Comparing two means (independent samples)

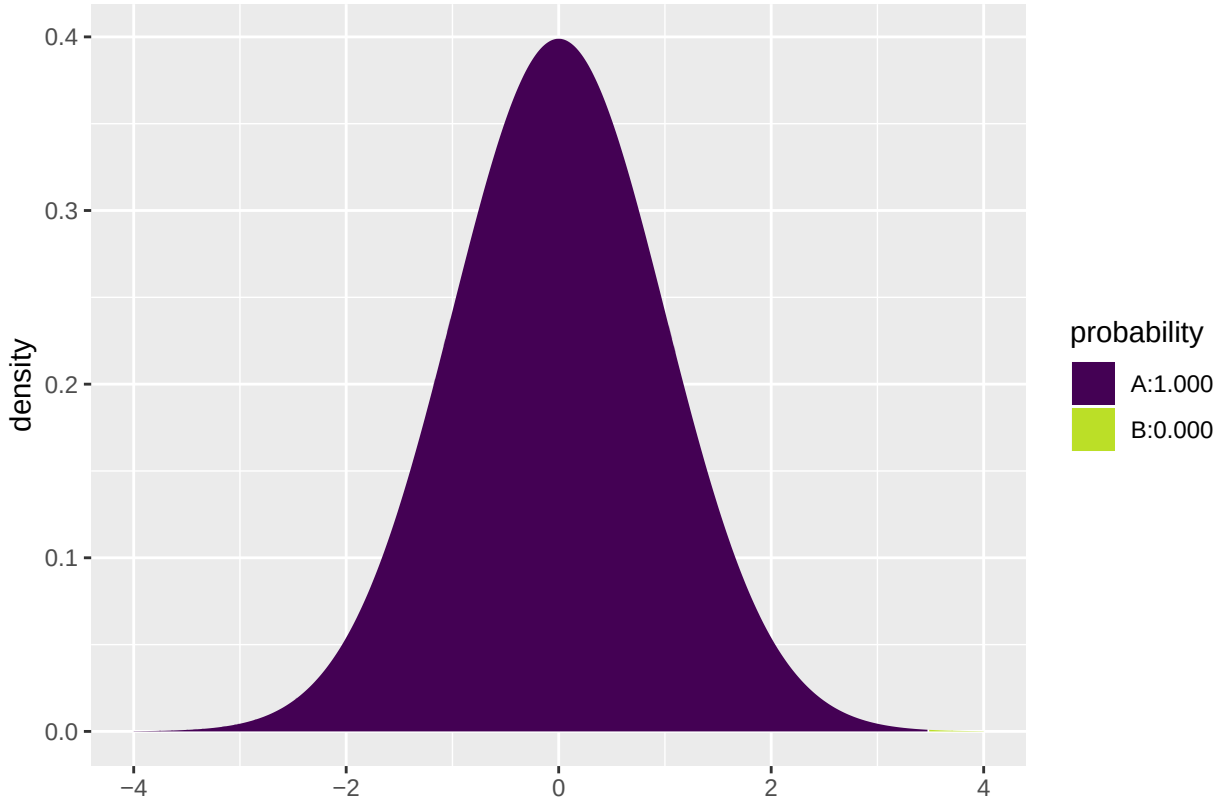
We return to the `Chile` data. We study the association between the variables `sex` and `statusquo` (scale of support for the status-quo). So, we will perform a significance test to test for difference in the mean of `statusquo` for male and females.

```
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.

1.17 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%.

1.18 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 $d_1, d_2, \ldots, d_{32}$ are now considered as **ONE** sample from a population with mean μ.
  - Test the hypothesis $H_0 : \mu = 0$ as usual (using a t-test for testing the mean as in the previous lecture).

1.18.1 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
```

2 Comparison of two proportions

2.1 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 π_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 π_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.

2.2 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.

2.3 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 π, 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, π, 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.

2.4 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: $\hat{\pi}_2 = \frac{363}{363+946} = 0.277$

2.5 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:

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

2.6 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.

2.7 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
```

2.8 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 `voteNo` proportions for women and men.

2.9 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
Estimation			
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	$(\hat{\pi}_2 - \hat{\pi}_1) \pm z(se)$	$(\bar{y}_2 - \bar{y}_1) \pm t(se)$
Significance testing			
1.	Assumptions	Randomization ≥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)	