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

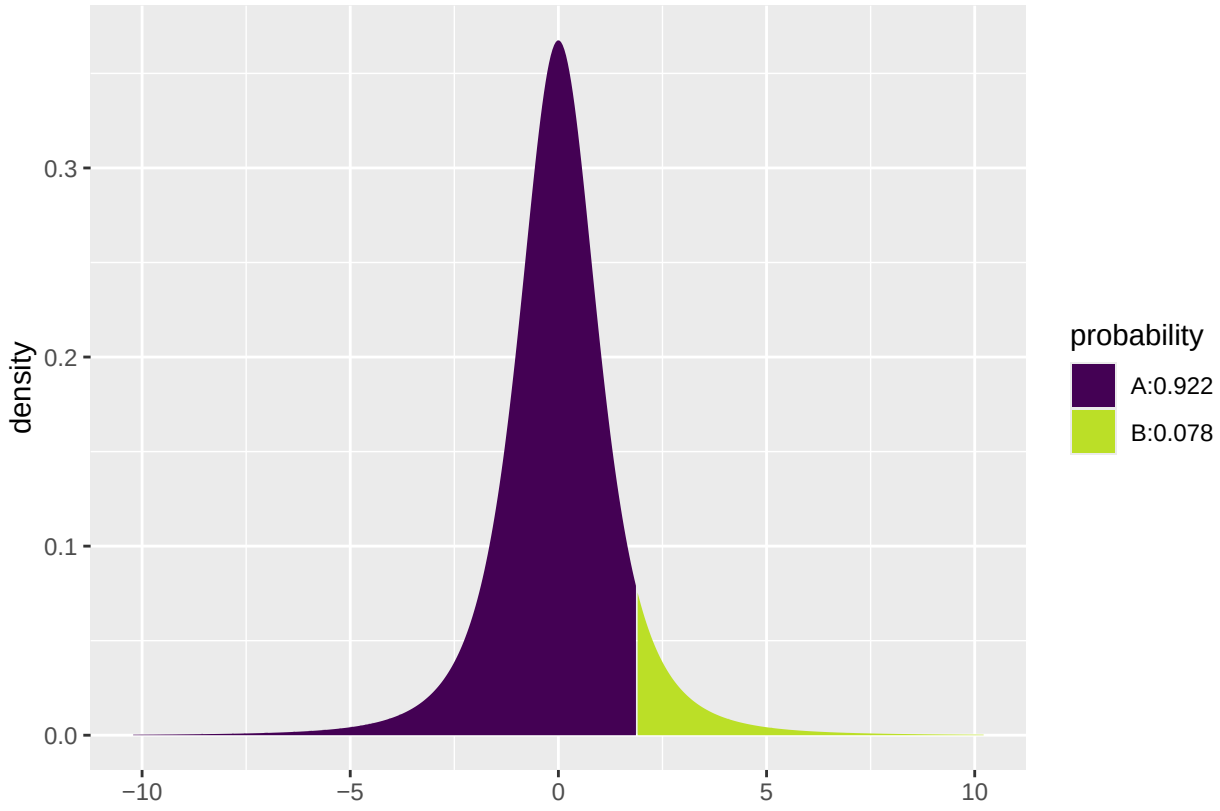
1.7.1 Two-sided t-test for mean:

- 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?

```r
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

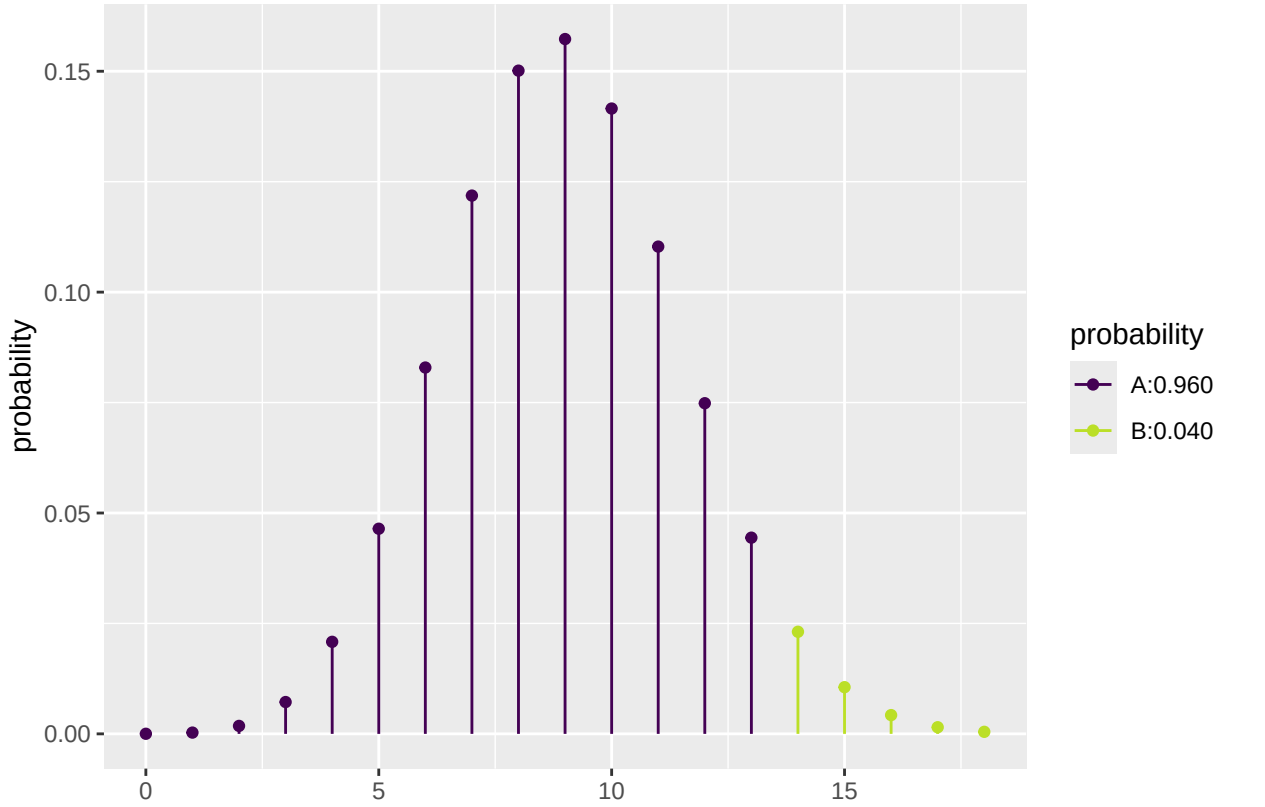
- Consider again a sample of size n, where we observe whether a given property is present or not.
- The relative frequency of the property in the population is π, which is estimated by $\hat{\pi}$.
- Let $y_+ = n\hat{\pi}$ be the frequency (total count) of the property in the sample.
- It can be shown 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

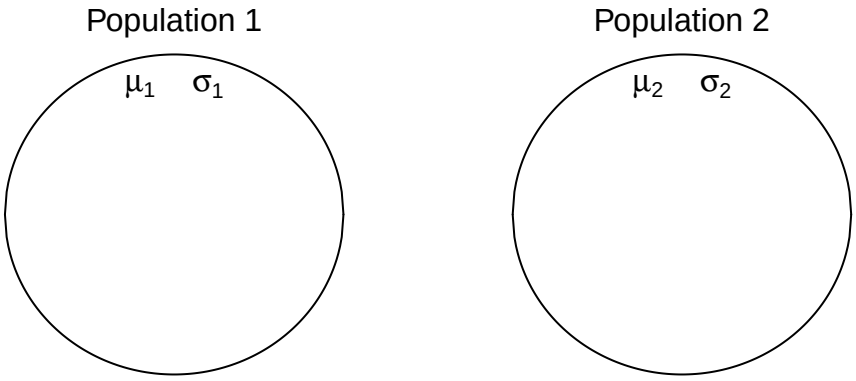
TABLE 6.7: Summary of Significance Tests for Means and Proportions

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	

2 Comparison of two populations

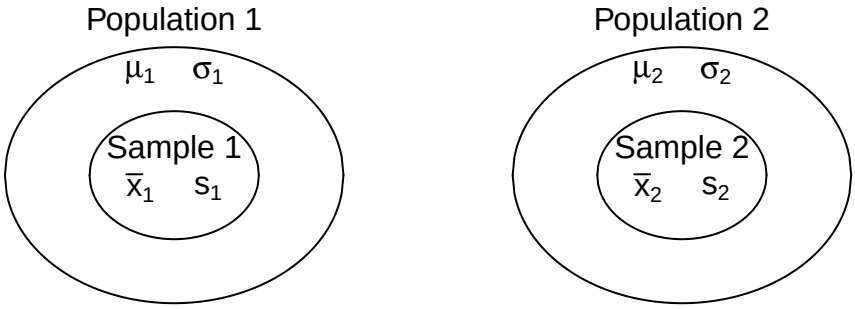
2.1 Two populations

- Consider two populations:
  - Population 1 has mean μ_1 and standard deviation σ_1.
  - Population 2 has mean μ_2 and standard deviation σ_2.
- We want to compare the means by looking at the difference $\mu_1 - \mu_2$.



2.2 Two samples

- We now take a sample from each population.
  - The sample from Population 1 has sample mean $\bar{x}_1$, sample standard deviation s_1 and sample size n_1.
  - The sample from Population 2 has sample mean $\bar{x}_2$, sample standard deviation s_2 and sample size n_2.



2.3 Dependent and independent samples

- We distinguish between two types of samples:
  - The two samples are **independent**.
  - The two samples are **paired**.
- **Example:** Suppose we consider the fuel consumption of cars.
  - If we compare two samples of cars with different engine types, then the two samples are *independent*, since each car can only have one of the two engine types.
  - If we compare the fuel consumption of cars at two different speed levels by testing each car at both speed levels, then then samples are *paired*.

2.4 Comparison of two means (Independent samples)

- We consider the situation, where we have two independent samples of a quantitative variable.
- We estimate the difference $\mu_1 - \mu_2$ by
$$d = \bar{x}_1 - \bar{x}_2.$$
- Assume that we can find the **estimated standard error** se_d of the difference.

- If the samples come from two normal distributions, or if both samples are large ($n_1, n_2 \geq 30$), then one can show
$$T_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{se_d} \sim \mathtt{t}(df),$$
where $\mathtt{t}(df)$ is a t-distribution with df degrees of freedom.
- By the usual procedure, we can use this to construct a confidence interval for the unknown population difference of means $\mu_1 - \mu_2$ by
$$(\bar{x}_1 - \bar{x}_2) \pm t_{crit} se_d,$$
where the critical t-score, t_{crit}, is determined by the confidence level and the df.

2.5 Significance test (Independent samples)

- We may be interested the testing the **null-hypothesis** that the population means are the same, which we can formulated as:
  - $H_0 : \mu_1 - \mu_2 = 0$.
  - $H_a : \mu_1 - \mu_2 \neq 0$.
- If the null hypothesis is true, then the **test statistic**:
$$T_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{se_d},$$
has a t-distribution with df degrees of freedom.
- The **p-value** is the probability of observing something further away from 0 than t_{obs} in a $\mathtt{t}(df)$ distribution.
- It remains to find the estimated standard error se_d and the degrees of freedom df. We distinguish between two cases:
  - The two populations have equal variances $\sigma_1^2 = \sigma_2^2$.
  - The two populations have different variances $\sigma_1^2 \neq \sigma_2^2$.

2.6 Standard error (Independent samples, equal variances)

- The standard error of $d = \bar{x}_1 - \bar{x}_2$ is given by the formula:
$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$
- If the **variances are equal**, $\sigma_1^2 = \sigma_2^2$, then we estimate the common value by the **pooled variance estimate**
$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$
- Inserting this estimate in the formula for the standard error we obtain the estimated standard error
$$se_d = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$
- In this situation, the degrees of freedom are $df = n_1 + n_2 - 2$.

2.7 Example: Comparing two means (independent samples, equal variances)

We return to the `mtcars` data. We study the association between the variables `vs` and `mpg` (engine type and fuel consumption). So, we will perform a significance test to test the null-hypothesis that there is no difference between the mean of fuel consumption for the two engine types.

- We will test the null-hypothesis assuming equal variances:

```
library(mosaic)
fv <- favstats(mpg ~ vs, data = mtcars)
fv
```

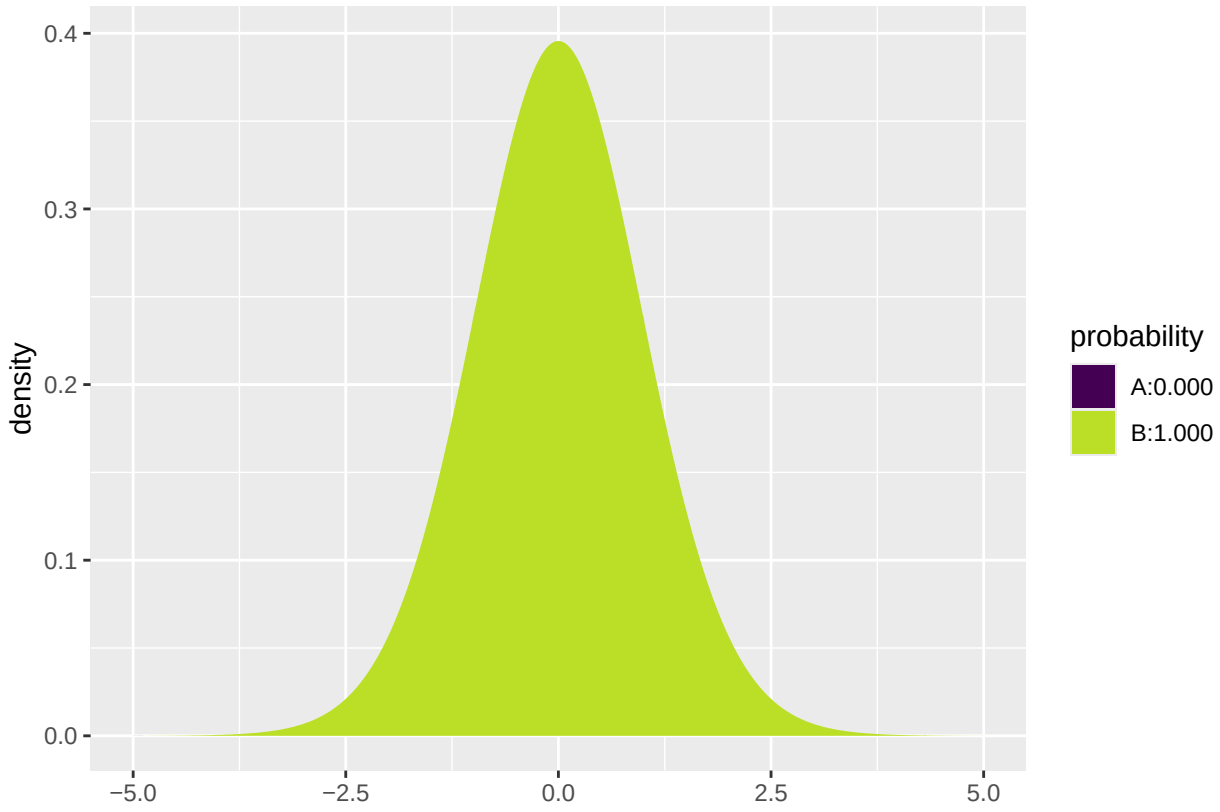
```
##   vs  min   Q1 median   Q3  max mean   sd  n missing
## 1  0 10.4 14.8   15.7 19.1 26.0 16.6 3.86 18       0
## 2  1 17.8 21.4   22.8 29.6 33.9 24.6 5.38 14       0
```

- Difference: $d = 16.6167 - (24.5571) = -7.9405$.
- Sample sizes: $n_1 = 18$ and $n_2 = 14$.
- Estimated standard deviations: $s_1 = 3.8607$ (not v-shaped) and $s_2 = 5.379$ (v-shaped).
- Pooled variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{17 \cdot 3.8607^2 + 13 \cdot 5.379^2}{18+14-2} = 20.984.$$

- Estimated standard error of difference: $se_d = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = \sqrt{20.984}\sqrt{\frac{1}{18}+\frac{1}{14}} = 1.6324$.
- Observed t-score for $H_0: \mu_1 - \mu_2 = 0$ is: $t_{obs} = \frac{d-0}{se_d} = \frac{-7.9405}{1.6324} = -4.864$.
- The degrees of freedom are $df = n_1 + n_2 - 2 = 30$.
- We find the p-value:

```
2*pdist("t", q = -4.864, df=30, xlim = c(-5, 5))
```




```
## [1] 3.419648e-05
```

2.8 Standard error (Independent samples, unequal variances)

- If the **variances are unequal**, then we simply insert the two estimates s_1^2 and s_2^2 for σ_1^2 and σ_2^2 in the formula for the standard error to obtain the estimated standard error

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

- The degrees of freedom df for se_d can be estimated by a complicated formula, which we will not present here (see p.365 in the book).
- Note:
  - If both n_1 and n_2 are above 30, then we may use the standard normal distribution to compute a z-score rather than the t-distribution to compute the t-score. This way we avoid computing df.
  - If n_1 or n_2 are below 30, then we let **R** calculate the degrees of freedom and the p-value/confidence interval.

2.9 Example: Comparing two means (independent samples, unequal variances)

We return to the `mtcars` data. We study the association between the variables `vs` and `mpg` (engine type and fuel consumption). So, we will perform a significance test to test the null-hypothesis that there is no difference between the mean of fuel consumption for the two engine types.

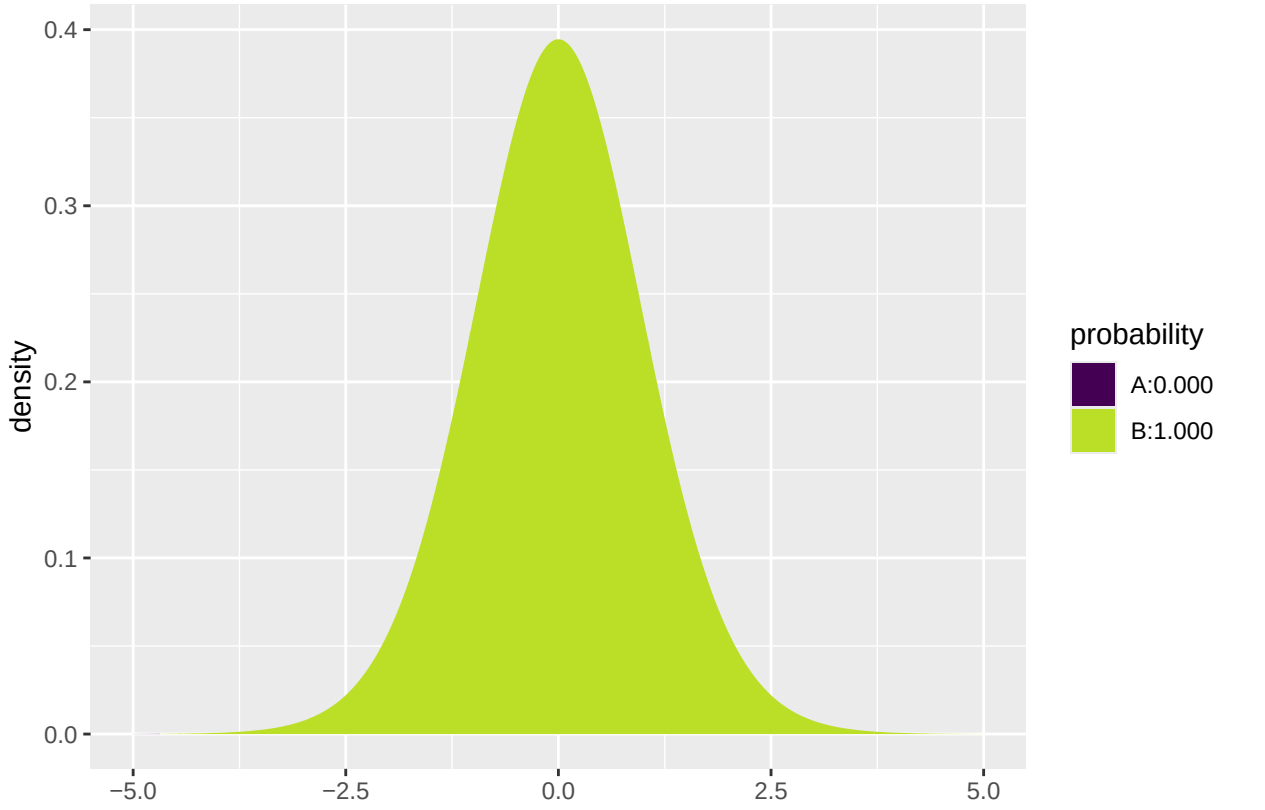
- We now make the analysis without assuming equal variances:

```
library(mosaic)
fv <- favstats(mpg ~ vs, data = mtcars)
fv
```

```
##   vs  min   Q1 median   Q3  max mean   sd  n missing
## 1  0 10.4 14.8   15.7 19.1 26.0 16.6 3.86 18       0
## 2  1 17.8 21.4   22.8 29.6 33.9 24.6 5.38 14       0
```

- Difference: $d = 16.6167 - (24.5571) = -7.9405$.
- Sample sizes: $n_1 = 18$ and $n_2 = 14$.
- Estimated standard deviations: $s_1 = 3.8607$ (not v-shaped) and $s_2 = 5.379$ (v-shaped).
- Estimated standard error of difference: $se_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3.8607^2}{18} + \frac{5.379^2}{14}} = 1.7014$.
- Observed t-score for $H_0: \mu_1 - \mu_2 = 0$ is: $t_{obs} = \frac{d-0}{se_d} = \frac{-7.9405}{1.7014} = -4.6671$.
- The degrees of freedom can be found using R (see below) to be $df = 22.716$.
- We find the p-value:

```
2* pdist("t", q = -4.6671, df=22.716, xlim = c(-5, 5))
```




```
## [1] 0.0001098212
```

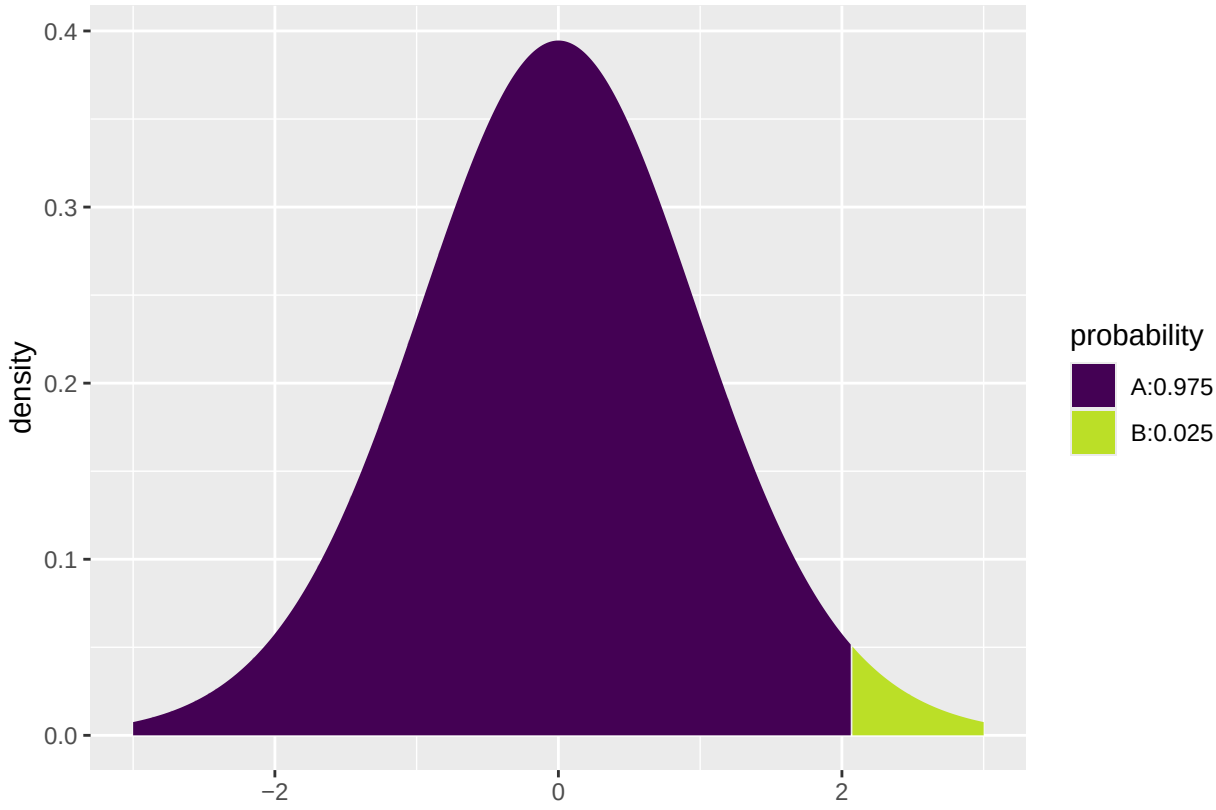
- We reject the null-hypothesis and conclude that the fuel consumption is different for the two engine types.

2.10 Example: Comparing two means (independent samples)

- Now we know there is a difference between the two population means. We can also make a 95% confidence interval for how large the difference $\mu_1 - \mu_2$ actually is by the formula

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

```r
qdist("t", p = 1-0.05/2, df=22.716, xlim = c(-3, 3))
```




```
## [1] 2.07009
```

- Inserting the values from the previous slide yields

$$[-7.94 - 2.07 * 1.70; -7.94 + 2.07 * 1.70] = [-11.5, -4.4].$$

- We are 95% confident that the difference in fuel consumption is between the two engine types is between -4.4mpg and -11.5mpg.

2.11 T-test in R (Independent samples)

- We can leave all the calculations to **R** by using `t.test`:

```r
t.test(mpg ~ vs, data = mtcars,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by vs
## t = -4.6671, df = 22.716, p-value = 0.0001098
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.462508  -4.418445
## sample estimates:
## mean in group 0 mean in group 1
##        16.61667        24.55714
```

- We recognize the t-score -4.6671, the p-value 0.0001, and the confidence interval $[-11.5; -4.4]$. The estimated degrees of freedom can be found in the output to be $df = 22.716$.

2.12 Test for equal variances (Independent samples)

- In order to decide whether to use the t-test with equal or unequal variance, we may test the hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$.
- As test statistic we use

$$F_{obs} = \frac{s_1^2}{s_2^2}.$$

- If the null-hypothesis is true, we expect F_{obs} to take values close to 1. Small and large values are critical for H_0.
- Under H_0, F_{obs} follows a so-called F-distribution with $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$ degrees of freedom.
  - If $F_{obs} < 1$ we reject the null-hypothesis if two times the probability of getting something smaller than F_{obs} is less than the significance level.
  - If $F_{obs} > 1$ we reject the null-hypothesis if two times the probability of getting something larger than F_{obs} is less than the significance level.

2.12.1 Example: Test for equal variances (Independent samples)

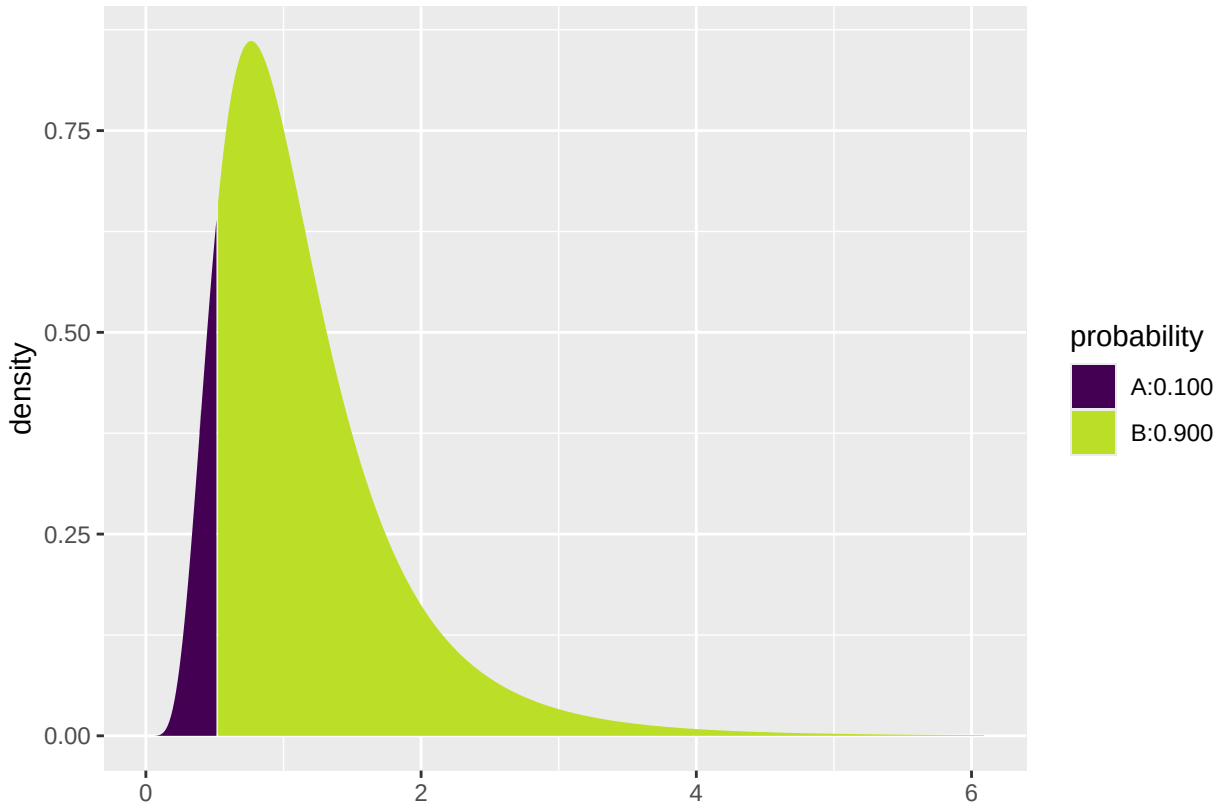
- To test whether the variance is the same for the two engine types in the `mtcars` example, we first compute the sample variances.

```
var(mpg~vs,data=mtcars)
```

```
##        0        1
## 14.90500 28.93341
```

- We compute $F_{obs} = \frac{s_1^2}{s_2^2} = \frac{14.9}{28.9} = 0.516$.
- The probability of observing something smaller than F_{obs} in an F-distribution with $df_1 = n_1 - 1 = 17$ and $df_2 = n_2 - 1 = 13$:

```
pdist("f", 0.516, df1=17, df2=13)
```




```
## [1] 0.1004094
```

- The p-value is $2 * 0.1004 = 0.2008$. Here we multiply by two because the test is two-sided (large values would also have been critical).
- We find no evidence against the null-hypothesis.

2.13 Comparison of two means: paired t-test (dependent samples)

- We now consider the case where we have two samples from two different populations but the observations in the two samples are **paired**.
  - For each pair, we can compute the difference between the two observations.
  - We now have one sample of observed differences.
  - We apply the the one-sample t-test from Lecture 2.1 to test whether the mean difference is zero.
- **Example:** Suppose we make the following experiment:
  - 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.
  - We are interested in 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 **paired** 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 one-sample t-test).

2.13.1 Reaction time: data 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
```

- We first 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")
all(yes$student == no$student)
```

```
## [1] TRUE
```

```
reaction_paired <- data.frame(student = no$student, yes = yes$reaction_time, no = no$reaction_time)
reaction_paired$diff <- reaction_paired$yes - reaction_paired$no
head(reaction_paired)
```

```
##   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_paired)
```

```
##
##  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
```

- With a p-value of 0.0000058 we reject the null-hypothesis that speaking on the phone has no influence on the reaction time.
- We can avoid the manual calculations and let **R** perform the significance test by using `t.test` with `paired = TRUE`:

```r
t.test(reaction_paired$no, reaction_paired$yes, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  reaction_paired$no and reaction_paired$yes
## t = -5.4563, df = 31, p-value = 5.803e-06
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -69.54814 -31.70186
## sample estimates:
## mean difference
##         -50.625
```

2.14 Response variable and explanatory variable

- The situation with two populations is an example where we have: * A **response variable** (or outcome, dependent variable).
  - An **explanatory variable** (or independent variable, covariate) that divides data in 2 groups.
- We are interested in the effect of the explanatory variable on the response variable.
  - For instance in the `mtcars` data, `mpg` is the response variable and `vs` is the explanatory variable.
- In this lecture we consider the case with one discrete explanatory variable. Module 3 is concerned with the case of one or more continuous variables.

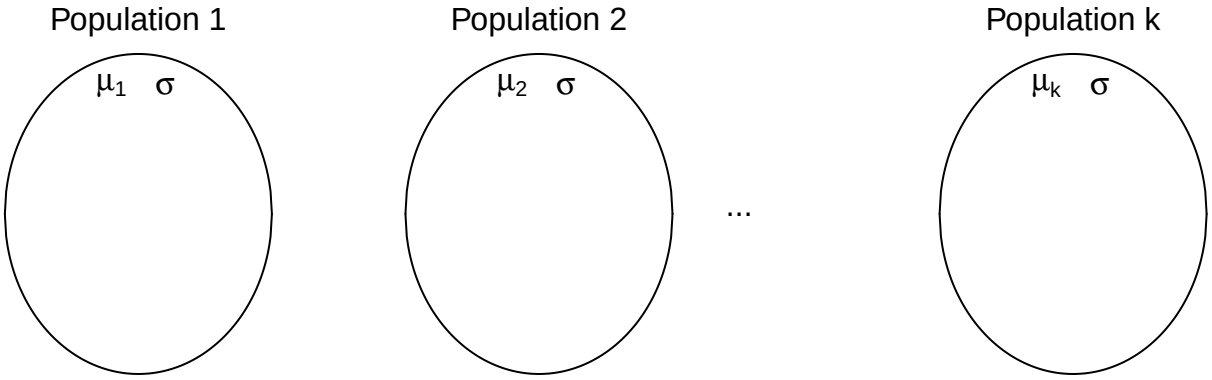
3 More than two groups (Analysis of variance)

3.1 More than two populations

- We are now going to consider a situation where we have k populations with mean values $\mu_1, \ldots, \mu_k$.
- We assume that each population follows a normal distribution and that the standard deviation is the same in all populations.
- We are interested in the null-hypothesis that all k populations have the same mean, i.e.

$$H_0 : \mu_1 = \cdots = \mu_k.$$
$$H_a : \text{ not all } \mu_1, \ldots \mu_k \text{ are the same.}$$

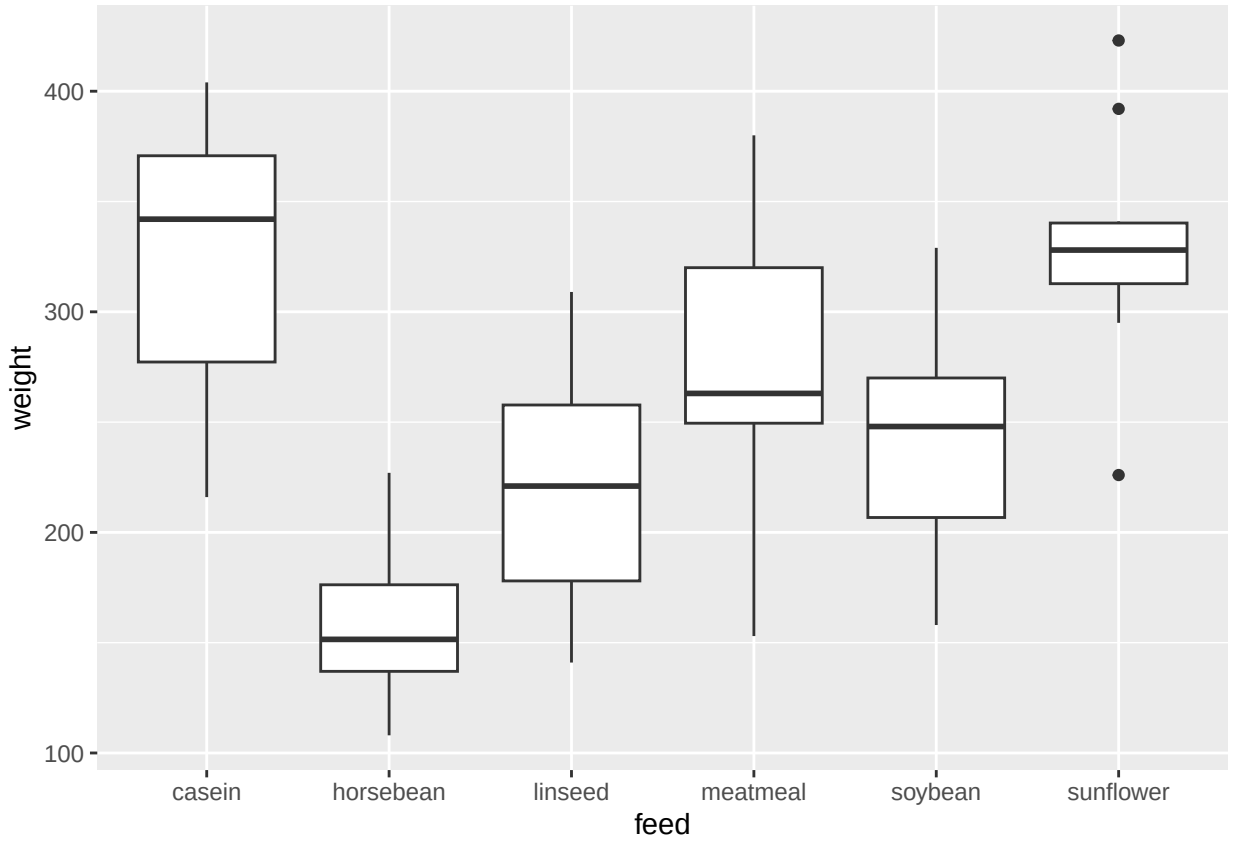
- We take out a sample from each population.



3.1.1 Data example

- The data set `chickwts` is available in `R`, and on the course webpage.
- 71 newly hatched chickens were randomly allocated into six groups, and each group was given a different feed supplement.
- Their weights in grams after six weeks are given along with feed types, i.e. we have a sample with corresponding measurements of 2 variables:
  - `weight`: a numeric variable giving the chicken weight.
  - `feed`: a factor giving the feed type.
- Always start with some graphics:

```
library(mosaic)
gf_boxplot(weight ~ feed, data = chickwts)
```



3.2 Estimation of mean values

- We estimate the mean in each group by the sample mean inside that group, i.e. $\hat{\mu}_i = \bar{x}_i$, $i = 1, \ldots, k$.
- We use `mean` to find the mean, for each group:

```
mean(weight ~ feed, data = chickwts)
```

```
##    casein horsebean   linseed  meatmeal   soybean sunflower
##  323.5833  160.2000  218.7500  276.9091  246.4286  328.9167
```

- We can e.g. see that the sample mean is 323.6, when `feed=casein` but 160.2, when `feed=horsebean`.
- Is this a significant difference ?

3.3 Contrasts

- If we want compare groups, it is convenient to formulate the model using contrasts.
- One group is chosen as the **reference group**, which all other groups are compared to.
  - Sometimes there is a group corresponding to "no treatment" and we are interested in the effect of different treatments. Other times the reference group can be arbitrary.
- If group 1 is the reference group, the mean values in the remaining groups groups can be expressed as

$$\mu_i = \mu_1 + \alpha_i,$$

where $\alpha_i = (\mu_i - \mu_1)$ is the difference between group i and the reference group. The α_i are called **contrasts**.

3.3.1 Example: contrast estimates

```
model <- lm(weight ~ feed, data = chickwts)
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

- In the example the groups are different `feeds`. R chooses the lexicographically smallest, which is `casein`, to be the reference group.
- We get information about contrasts and their significance:
- `Intercept` is the estimated mean $\hat{\mu}_{casein} = 323.583$ in the reference group.
  - In the same line, there is also a test of the null-hypothesis $H_0 : \mu_1 = 0$ that the weight after 6 weeks is 0 ($p < 2 \times 10^{-16}$) (of course, chickens grow a lot over 6 weeks).
- The line `feedhorsebean` estimates the contrast $\alpha_{horsebean}$ between the `casein` and `horsebean` group to be $\hat{\alpha}_{horsebean} = -163.383$.
  - The null-hypothesis that there is no difference between casein and horsebean ($H_0 : \alpha_{horsebean} = 0$) is rejected with p=$2 \times 10^{-9}$.

3.4 Overall test for effect

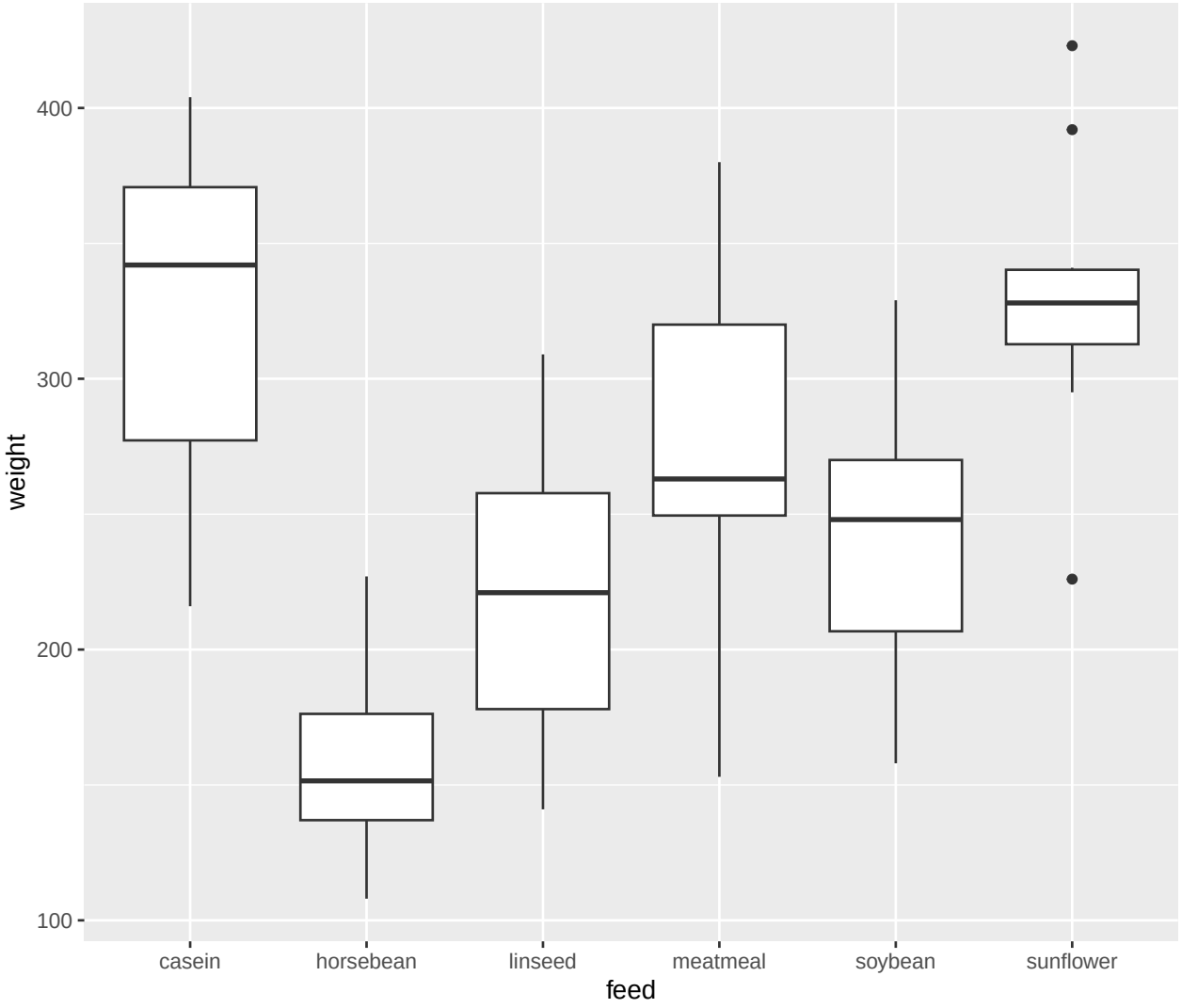
- We are now interested in testing the null-hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \quad \text{against} \quad H_a : \text{ Not all of the population means are the same}$$

- Alternatively

$$H_0 : \alpha_2 = \alpha_3 = \cdots = \alpha_k = 0, \quad H_a : \text{ At least one contrast is non-zero.}$$

- Idea: Compare variation within groups and variation between groups.



3.5 Test statistic

- We use the test statistic

$$F_{obs} = \frac{(TSS - SSE)/(k-1)}{SSE/(n-k)}.$$

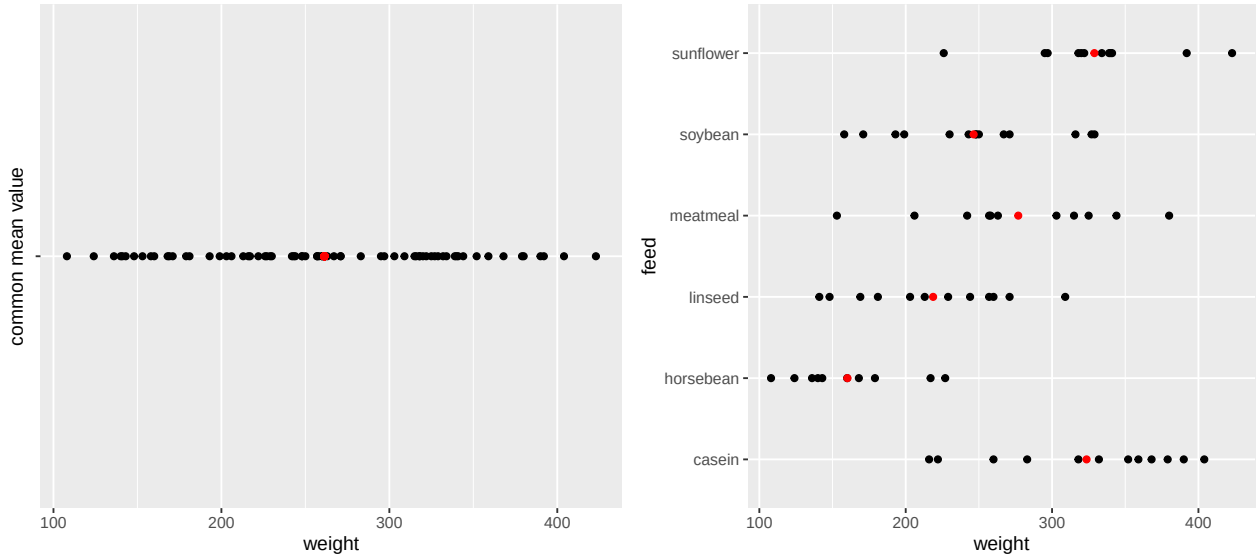
- If observations from group i are called x_{ij}, $j = 1, \ldots, k$, we have:
  - $TSS = \sum_i \sum_j (x_{ij} - \bar{x})^2$, where $\bar{x}$ is the average of all observations from all groups.

  - $SSE = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2$.

- Interpretation:
  - TSS: error sum of squares if common mean.
  - SSE: error sum of squares if different means.
  - TSS-SSE: how much does error sum of squares increase if means are restricted to be equal.
- One can show that TSS-SSE measures the variance of group means around common mean.
- Interpretation is thus

$$F_{obs} = \frac{\text{variance between groups}}{\text{variance within groups}}.$$

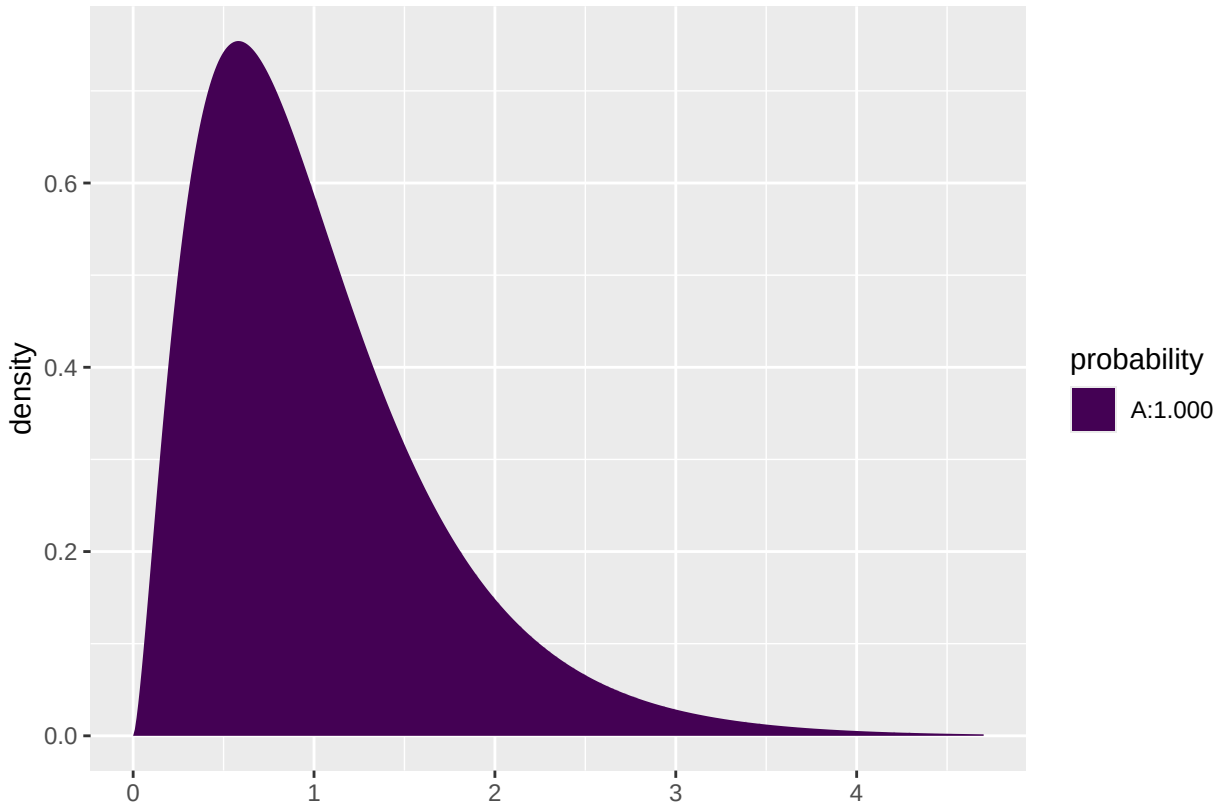
```
## Warning in geom_point(aes(x = red_dot), color = "red"): All aesthetics have length 1, but the data ha
## i Please consider using `annotate()` or provide this layer with data containing a single row.
```



3.6 The F-test

- A large variation between groups compared to the variation within groups points against H_0.
- Thus, large values are critical for the null-hypothesis.
- Under the null-hypothesis, F_{obs} follows an F-distribution with $df_1 = k - 1$ and $df_2 = n - k$ degrees of freedom.
- A p-value for the null-hypothesis is the probability of observing something larger than F_{obs} in an F-distribution with df_1 and df_2 degrees of freedom.
- For instance if $F_{obs} = 15.36$ with $df_1 = 5$ and $df_2 = 65$ degrees of freedom:

```
1 - pdist("f", 15.36, df1=5, df2=65)
```




```
## [1] 5.967948e-10
```

3.7 Example

```
model <- lm(weight ~ feed, data = chickwts)
summary(model)
```

```
##
## Call:
## lm(formula = weight ~ feed, data = chickwts)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.909  -34.413    1.571   38.170  103.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    323.583     15.834  20.436  < 2e-16 ***
## feedhorsebean -163.383     23.485  -6.957 2.07e-09 ***
## feedlinseed   -104.833     22.393  -4.682 1.49e-05 ***
## feedmeatmeal   -46.674     22.896  -2.039 0.045567 *
## feedsoybean    -77.155     21.578  -3.576 0.000665 ***
## feedsunflower    5.333     22.393   0.238 0.812495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.85 on 65 degrees of freedom
## Multiple R-squared:  0.5417, Adjusted R-squared:  0.5064
```

```
## F-statistic: 15.36 on 5 and 65 DF,  p-value: 5.936e-10
```

- The last line gives us the value of $F_{obs} = 15.36$ and the corresponding p-value (5.9×10^{-10}). Clearly there is a significant difference between the types of `feed`.