Comparison of two or more groups

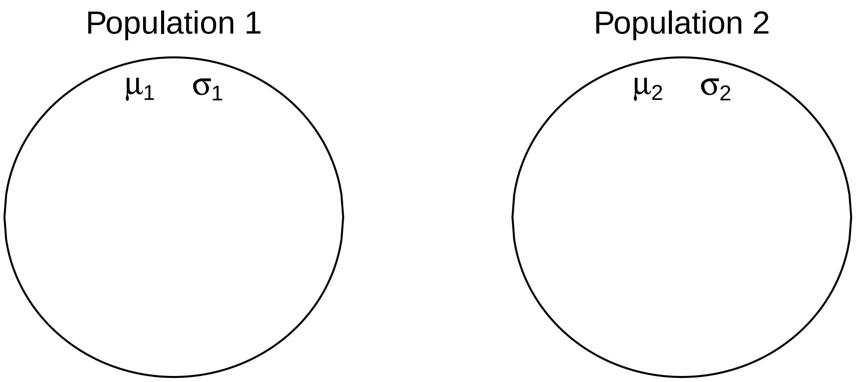
The ASTA team

Contents

1 Comparison of two populations

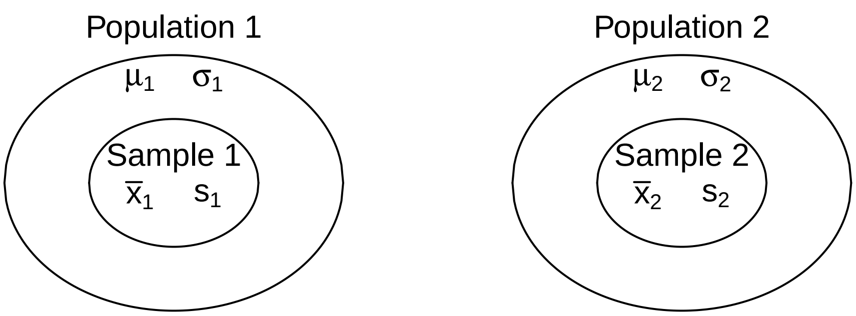
1.1 Two populations

- 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 $\mu_1 - \mu_2$.



1.2 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.



1.3 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*.

1.4 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.

1.5 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 $\sigma_1^2 \neq \sigma_2^2$.

1.6 Standard error (Independent samples, equal variances)

- The standard error of $d = \bar{x}_1 - \bar{x}_2$ is given by the formula:

$$\sqrt{\tfrac{\sigma_1^2}{n_1} + \tfrac{\sigma_2^2}{n_2}}.$$

- If the **variances are equal**, $\sigma_1^2 = \sigma_2^2$, then we estimate the common value by the **pooled variance estimate**

$$s_p^2 = \tfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

- Inserting this estimate in the formula for the standard error we obtain the estimated standard error

$$se_d = \sqrt{\tfrac{s_p^2}{n_1} + \tfrac{s_p^2}{n_2}} = s_p\sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}}.$$

- In this situation, the degrees of freedom are $df = n_1 + n_2 - 2$.

1.7 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
```

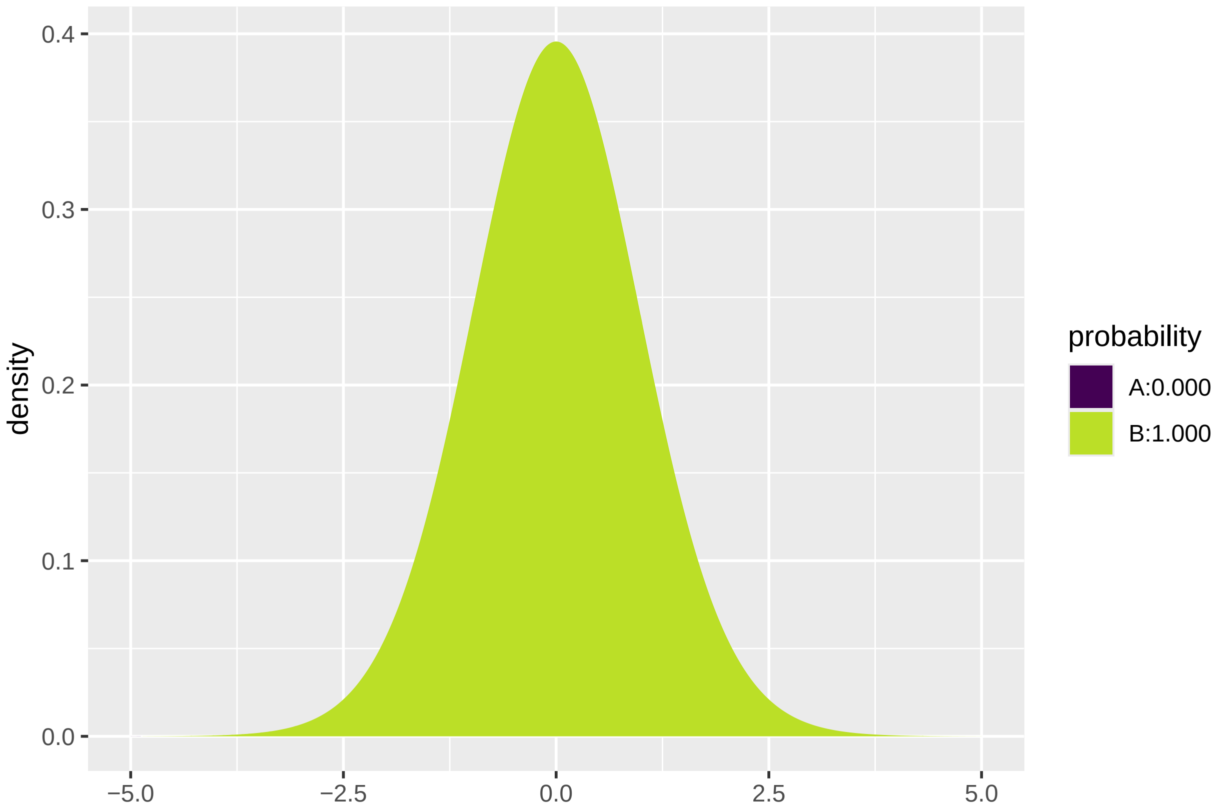
```
##   vs  min   Q1 median   Q3  max mean   sd  n missing
## 1  0 10.4 14.8   15.7 19.1 26.0 16.6 3.86 18       0
## 2  1 17.8 21.4   22.8 29.6 33.9 24.6 5.38 14       0
```

- 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
```

1.8 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.

1.9 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.

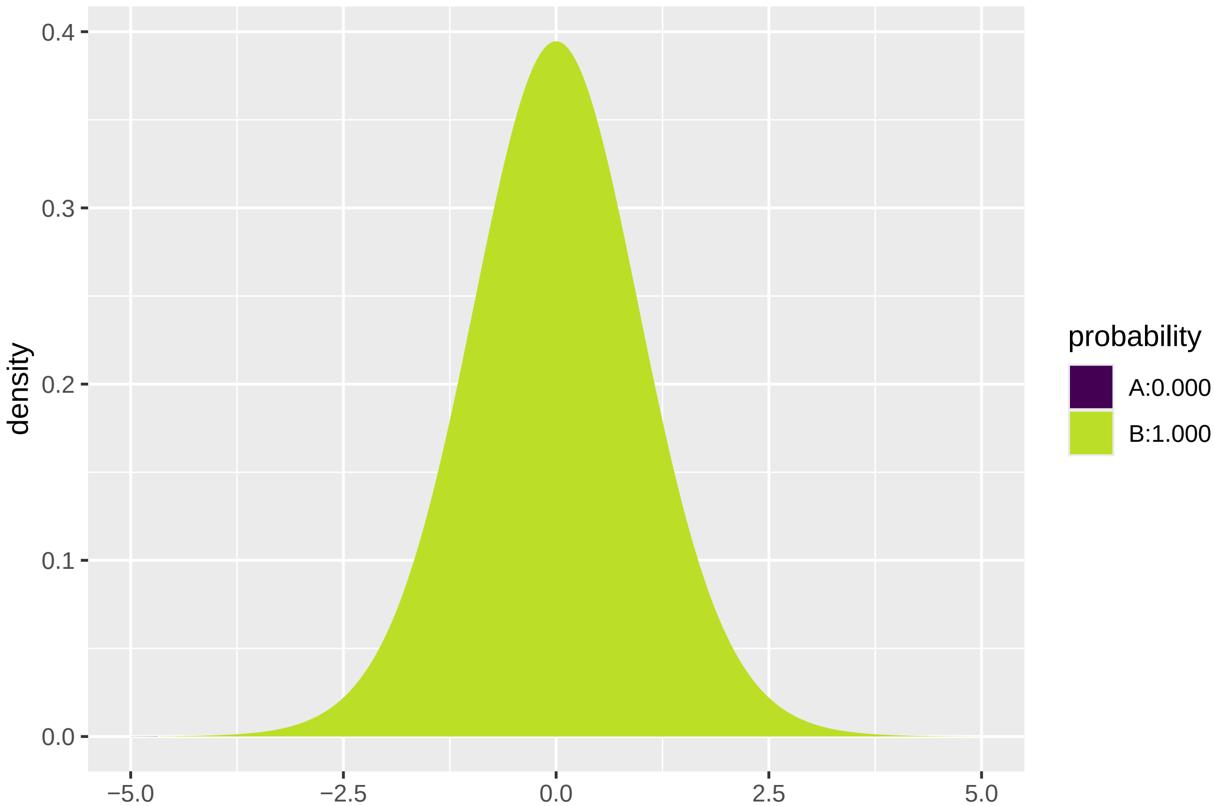
- We now make the analysis without assuming equal variances:

```
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
```

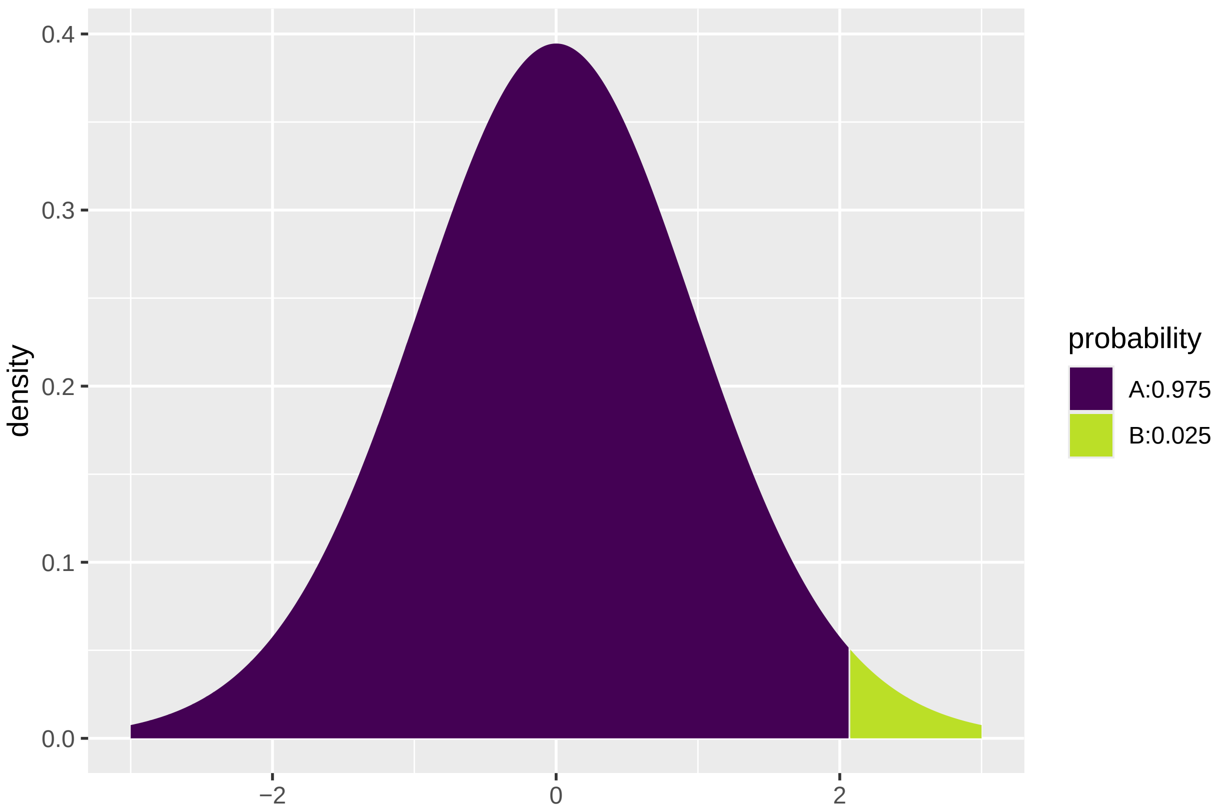
- We reject the null-hypothesis and conclude that the fuel consumption is different for the two engine types.

1.10 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.

1.11 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 $df = 22.716$.

1.12 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.

1.12.1 Example: Test for equal variances (Independent samples)

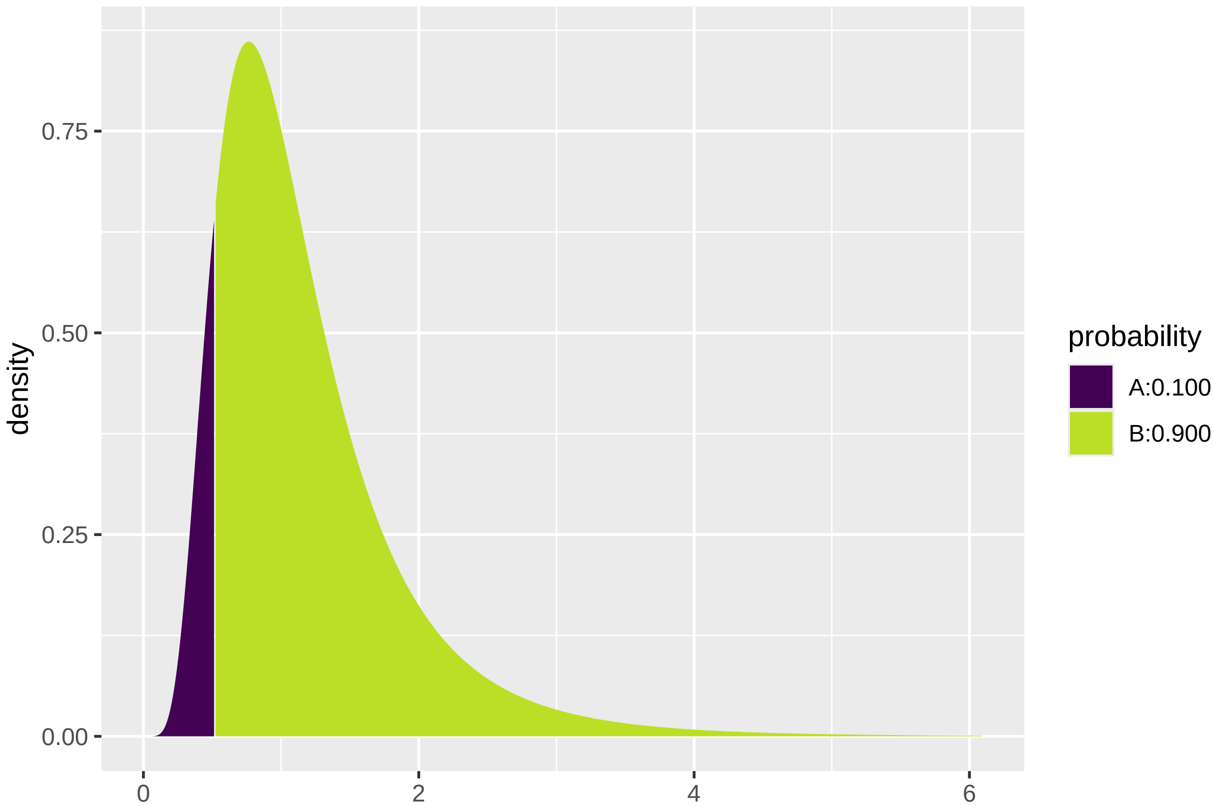
- To test whether the variance is the same for the two engine types in the `mtcars` example, we first compute the sample variances.

```
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.

1.13 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).

1.13.1 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
```

1.14 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.

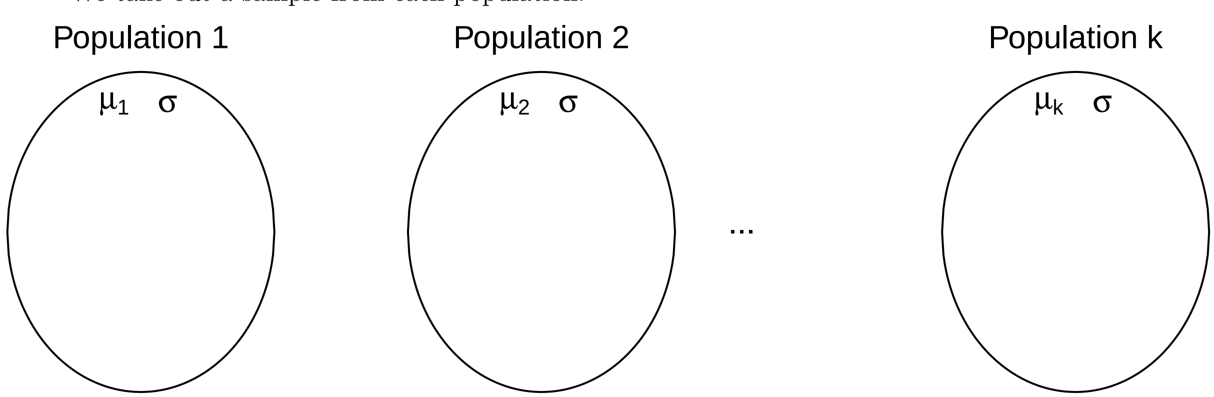
2 More than two groups (Analysis of variance)

2.1 More than two populations

- We are now going to consider a situation where we have k populations with mean values $\mu_1, \dots, \mu_k$.
- 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.

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

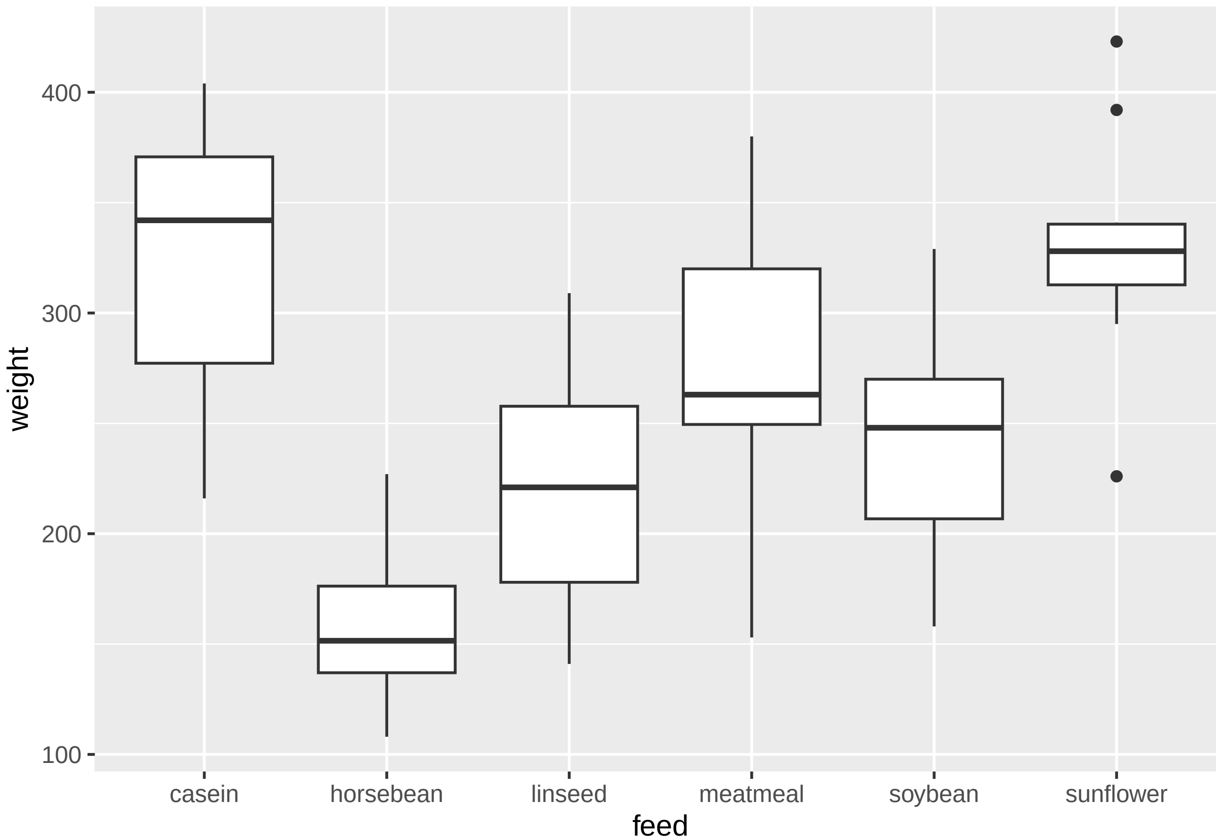
- We take out a sample from each population.



2.1.1 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)
```



2.2 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 ?

2.3 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**.

2.3.1 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=$2 \times 10^{-9}$.

2.4 Overall test for effect

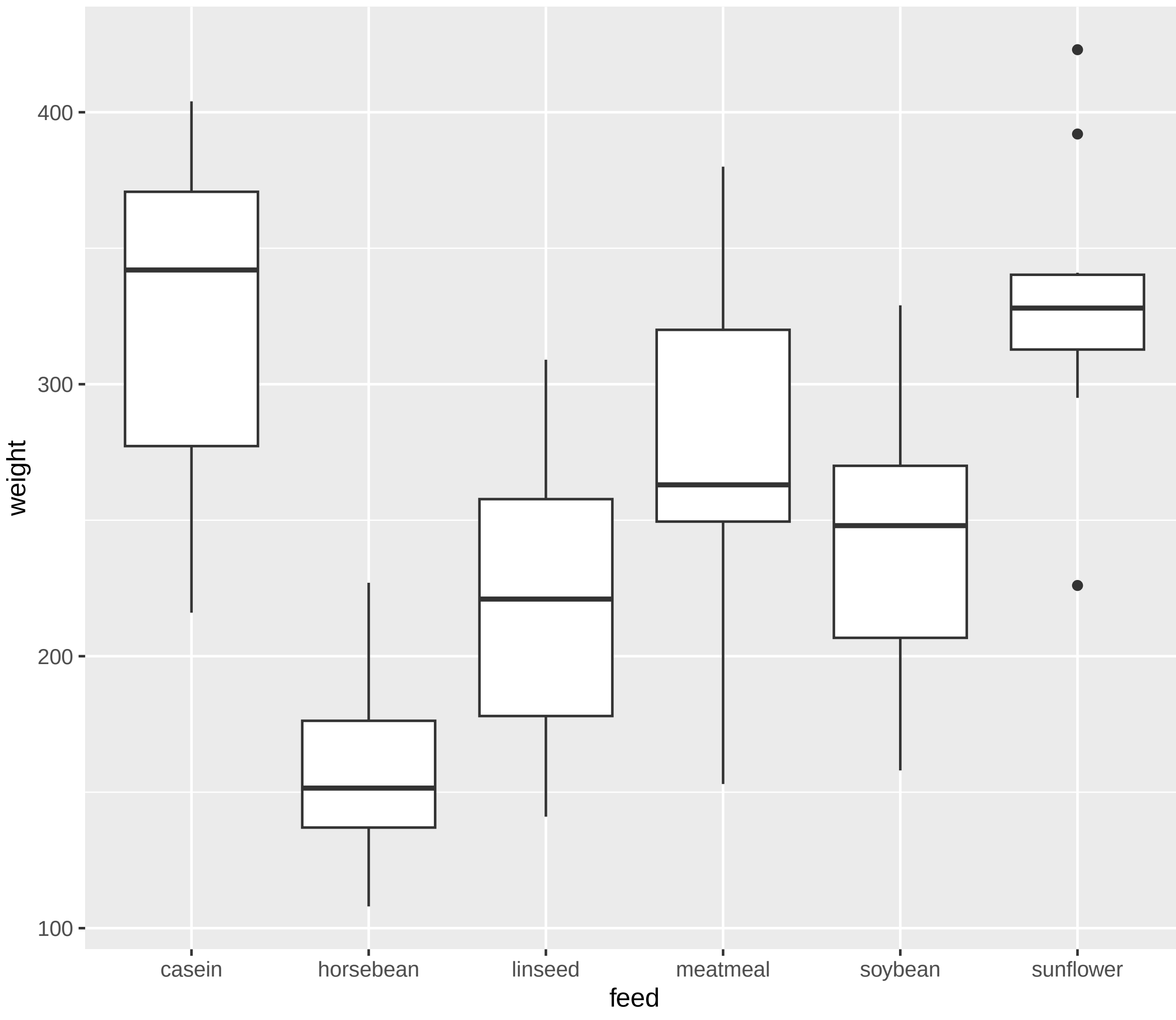
- We are now interested in testing the null-hypothesis

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

- 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.



2.5 Test statistic

- We use the test statistic

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

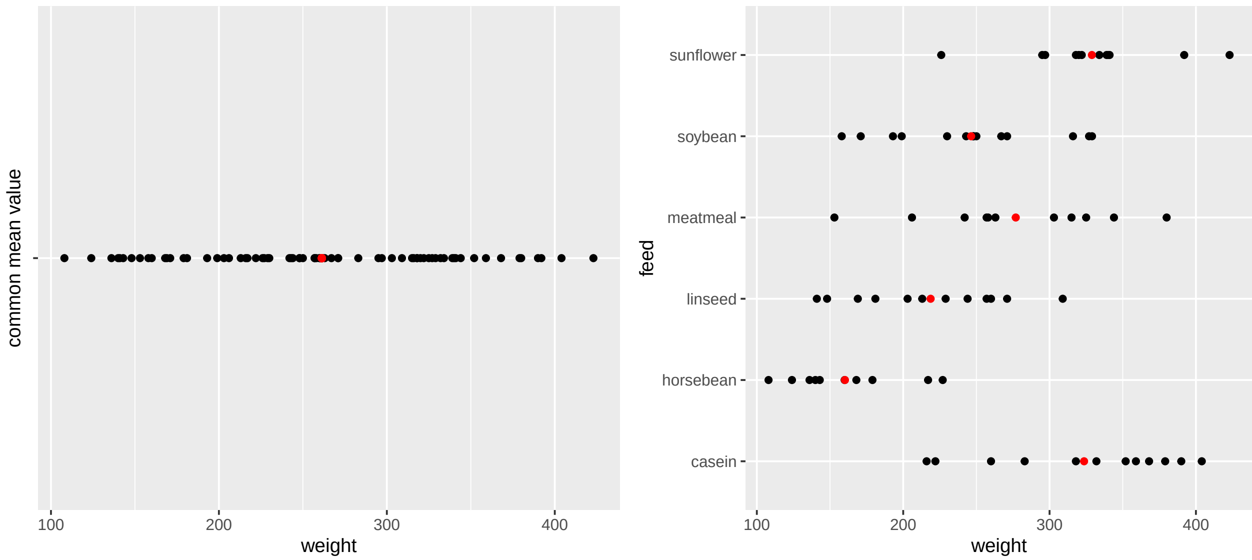
- If observations from group i are called x_{ij}, $j = 1, \ldots, k$, we have:
  - $TSS = \sum_i \sum_j (x_{ij} - \bar{x})^2$, where $\bar{x}$ is the average of all observations from all groups.

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

- 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

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

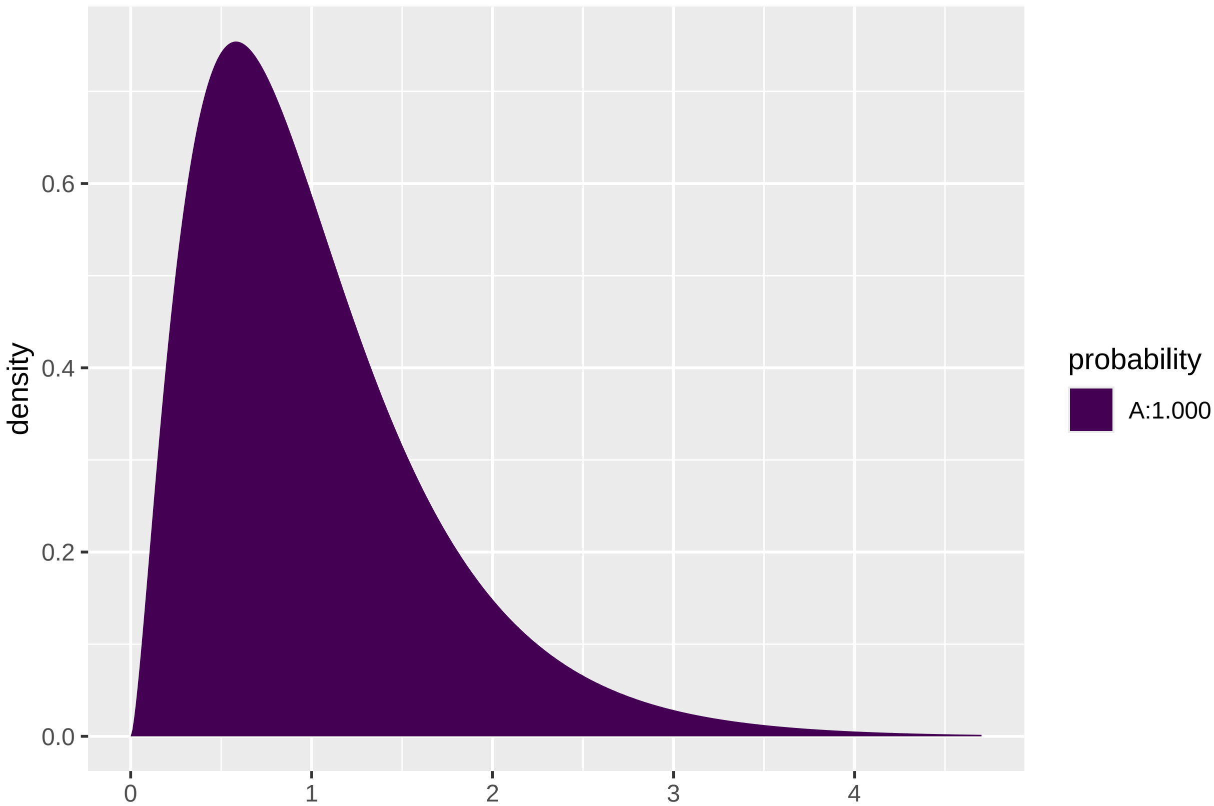
```
## Warning in geom_point(aes(x = red_dot), color = "red"): All aesthetics have length 1, but the data ha
## i Please consider using `annotate()` or provide this layer with data containing
##   a single row.
```



2.6 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
```

2.7 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`.