Two independent samples t-test

Danilo Sarti
danilo.sarti@mu.ie

https://danilosarti.github.io/

PRESS RECORD

Learning outcomes

- Know the motivation and when to use the t-test
- Explore a dataset to answer a question
- Know how to apply the test with R
- Extract conclusions from the test

Motivation

- Suppose we work for a brand that is producing a new drug for weight loss
- For some reason, you are asked to address the question: Do men and women differ in terms of weight on average? This will become the hypothesis to be tested (check previous classes!!)
- A statistical test can answer this question: the t-test (developed by Student)
- The t-test allows us to compare the means of two independent groups
- Interesting fact: The real Student was William Gosset (Guinness brewer!!)

Loading packages to help

- As we learned in the last lab sessions, let's load some R packages that will help us!!
  All code: https://github.com/danilosarti/t_test_class

```
library(tidyverse) #to help tidy data
library(ggpubr) #to produce some good graphs
library(rstatix) #to perform statistical tests in a easy manner!!!
```

The data.

- Suppose that your boss also provided you with a data set containing the weights of 20 men and 20 women.
- That data is stored in the R package datarium we used in previous classes. Let's take a look at the data.

```
# Load the data which is actually stored in the datarium R package.
data("genderweight", package = "datarium")
# Show a sample of the data by group
genderweight %>% sample_n_by(group, size = 2)
```

```
## # A tibble: 4 x 3
##   id    group weight
##   <fct> <fct>  <dbl>
## 1 14    F       64.1
## 2 9     F       62.9
## 3 38    M       81.6
## 4 21    M       89.7
```

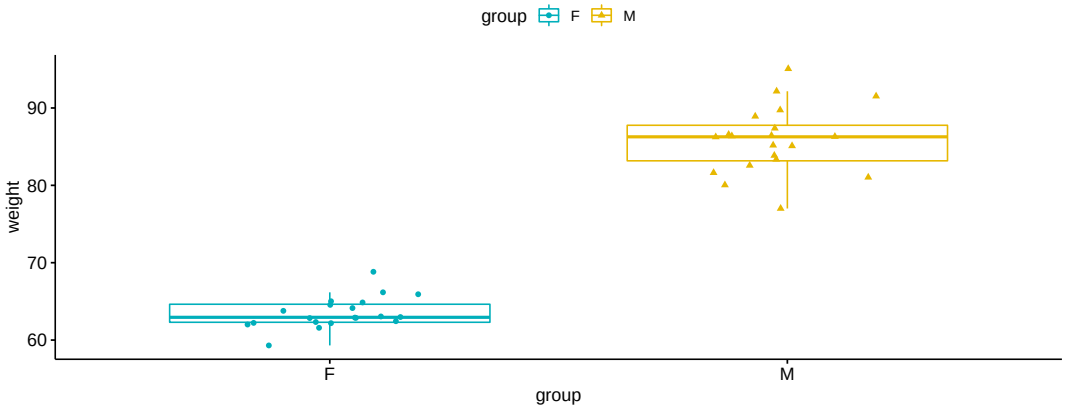
Data Visualisation

- As we learned is always helpful to visualize the data!!!
- Let us produce some boxplots with the code!!!

```
df=genderweight %>%
  group_by(group)
p=ggboxplot(df, x = "group", y = "weight",
                color = "group", palette =c("#00AFBB", "#E7B800"),
                add = "jitter", shape = "group")
```

What do we see in terms of averages?

```
p ## printing our nice plot!!
```



Data summary

- We can easily calculate some summary statistics
- Is there evidence that the means are different?
- How can we really test that? Solution: usage of a statistical test!!!

```
genderweight %>%
  group_by(group) %>%
  get_summary_stats(weight, type = "mean_sd")
```

```
## # A tibble: 2 x 5
##   group variable     n  mean    sd
##   <fct> <chr>    <dbl> <dbl> <dbl>
## 1 F     weight      20  63.5  2.03
## 2 M     weight      20  85.8  4.35
```

Recalling tests

- All tests have assumptions
- They are associated with something we call statistics (check previous material!!)
- They return a p-value associated with the hypothesis tested (H_0). If the p-value$<$0.05, or another value we may consider proper, it does not support the hypothesis
- They may have some additional features called parameters! (check the previous classes!!)

t-test and t-test

- We have, indeed two forms for the t-test
- The original Student test assumes that the variance (variability-sd^2) of the two groups are equal. Not so real for our data set!!
- The Welch version: Does not consider variances to be identical. It fits our case more!!
- Other names: independent t-test,independent samples t-test, unpaired t-test or unrelated t-test.
- The test considers the probability distribution of the differences between means of women and men

Proceedig the t-test in a pipe-friendly manner (tidyverse)

- Welch version of the test!! H_o =are the average weight of men and women different?
- We are using commands from the package rstatix
- We see The results above show the following components:
  .y.: the y variable used in the test – weight. group1,group2: the compared groups in the tests. statistic: Test statistic used to compute the p-value. df: degrees of freedom. A parameter associated to the test p: p-value. Is it less than 0.05?

```
stat.test <- genderweight %>%
  t_test(weight ~ group) %>%
  add_significance()
data.frame(stat.test) # to show complete results!
```

```
##      .y. group1 group2 n1 n2 statistic
## 1 weight      F      M 20 20 -20.79138
##         df       p p.signif
## 1 26.87236 4.3e-18     ****
```

Student t-test

- If you want to proceed with the original t-student test, use: `var.equal = TRUE`
- Do the final result regarding the means remains the same? H_o =are the average weight of men and women different?

```
stat.test2 <- genderweight %>%
  t_test(weight ~ group, var.equal = TRUE) %>%
  add_significance()
data.frame(stat.test) # to show complete results!
```

```
##      .y. group1 group2 n1 n2 statistic
## 1 weight      F      M 20 20 -20.79138
##         df       p p.signif
## 1 26.87236 4.3e-18     ****
```

How about the confidence interval for difference of the means?

- If you want to obtain also de confidence interval for difference of the means you can use the basic r function t.test!!

```
# Compute t-test
res <- t.test(weight ~ group, data = genderweight)
```

- conf.int is the confidence interval of the means difference at 95% (conf.int = [-24.5314, -20.1235])

```
res
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by group
## t = -20.791, df = 26.872, p-value <
## 2.2e-16
## alternative hypothesis: true difference in means between group F and gr
## 95 percent confidence interval:
##  -24.53135 -20.12353
## sample estimates:
## mean in group F mean in group M
##        63.49867        85.82612
```

We need to answer our boss!!

- Remember, your boss does not know data science. We need to inform him anyways!!
- The mean of the women is 63.5 with an associated standard deviation (sd) of 2.03; for men, the mean is 85.8(sd=4.3). A statistical test (Welch's t-test) returns a p-value <0.0001, which suggests that the weight of men and women differ on average. The results can also be visualized as follows:

```r
# Create a black and whit box-plot -- useful for some publications!!
bxp <- ggboxplot(
  genderweight, x = "group", y = "weight",
  ylab = "Weight", xlab = "Groups", add = "jitter"
  )

# Add p-value and significance levels obtained with code in previous slide
stat.test <- stat.test %>% add_xy_position(x = "group")
```

```
bxp +
  stat_pvalue_manual(stat.test, tip.length = 0) +
  labs(subtitle = get_test_label(stat.test, detailed = TRUE))
```