Introduction to Statistics with R

Session R04: t-Test

Marvin Schmitt

The example data set

This example data set contains depression measures before (`MZP1`) and after (`MZP2`) a novel therapy.

```
df = read.csv("R04_notes_dataset.csv")
nrow(df)
```

```
## [1] 35
```

```
colnames(df)
```

```
## [1] "ID"   "age"  "MZP1" "MZP2"
```

Data formats: `wide` and `long`

Data sets can come in `wide` format:

```
##   ID age MZP1 MZP2
## 1  1  43   37   29
## 2  2  23   19   16
## 3  3  21   23   15
```

or `long` format:

```
##   ID age  mzp depr
## 1  1  43 MZP1   37
## 2  1  43 MZP2   29
## 3  2  23 MZP1   19
## 4  2  23 MZP2   16
## 5  3  21 MZP1   23
## 6  3  21 MZP2   15
```

Data formats: conversion `wide` to `long`

```
df %>% head(3)
```

```
##   ID age MZP1 MZP2
## 1  1  43   37   29
## 2  2  23   19   16
## 3  3  21   23   15
```

```
df %>%
  gather(., mzp, depr, MZP1:MZP2) -> df_long

df_long %>% arrange(ID) %>% head(6)
```

```
##   ID age  mzp depr
## 1  1  43 MZP1   37
## 2  1  43 MZP2   29
## 3  2  23 MZP1   19
## 4  2  23 MZP2   16
## 5  3  21 MZP1   23
## 6  3  21 MZP2   15
```

Data formats: conversion `long` to `wide`

```
df_long %>% arrange(ID) %>% head(6)
```

```
##   ID age  mzp depr
## 1  1  43 MZP1   37
## 2  1  43 MZP2   29
## 3  2  23 MZP1   19
## 4  2  23 MZP2   16
## 5  3  21 MZP1   23
## 6  3  21 MZP2   15
```

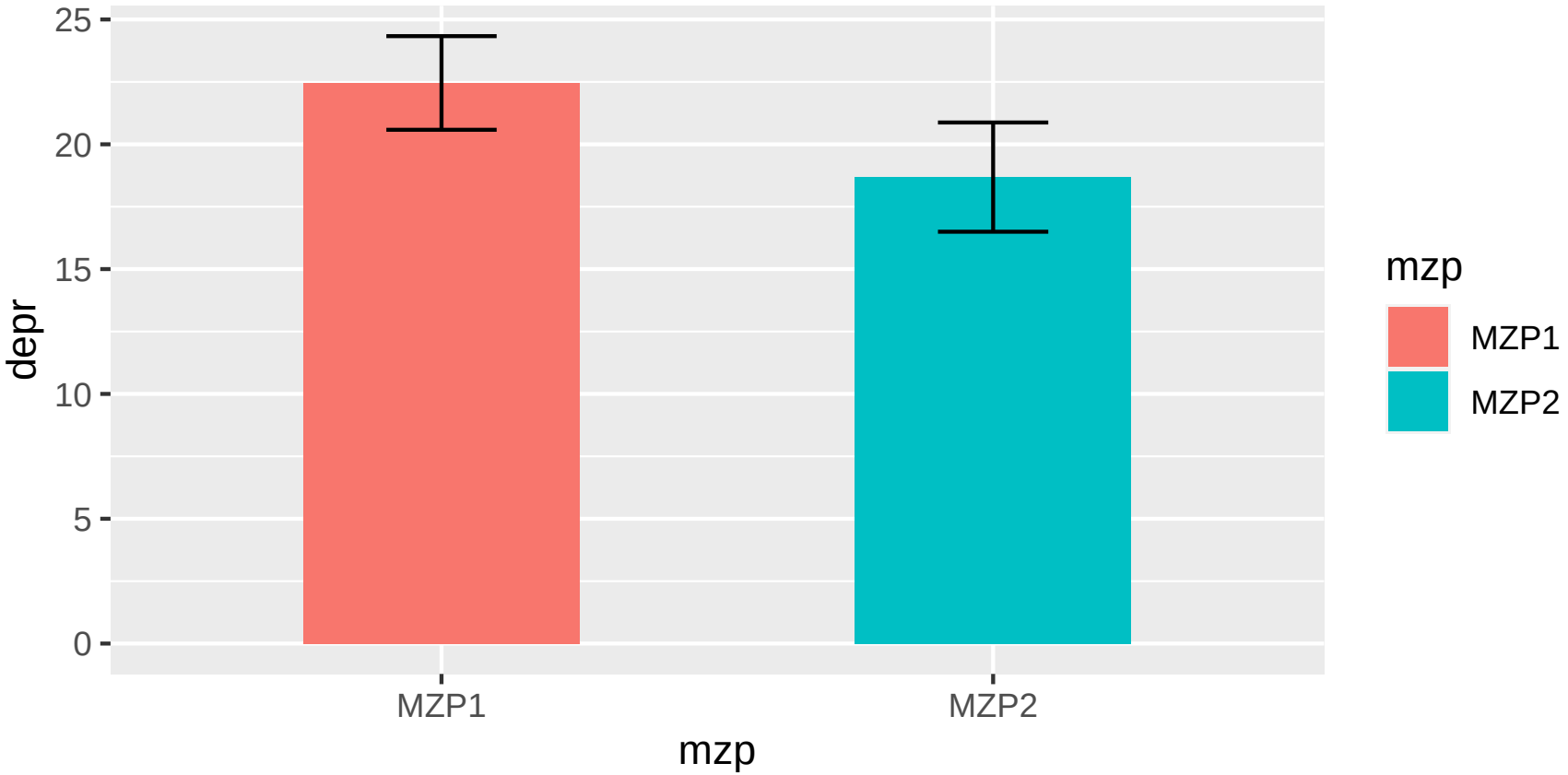
```
df_long %>%
  spread(., mzp, depr) -> df_wide

df_wide %>% head(3)
```

```
##   ID age MZP1 MZP2
## 1  1  43   37   29
## 2  2  23   19   16
## 3  3  21   23   15
```

ggplot2: Barplot with error bars

```
ggplot(data = df_long, aes(y=depr, x=mzp, fill=mzp)) +      # long format!
  stat_summary(fun = mean, geom = "bar", width=0.5) +
  stat_summary(fun.data = mean_se, geom = "errorbar", width=0.2)
```



t-test syntax

The function `t.test()` implements several versions of the $t-$test:

- One sample:
  - The first argument is the data, `mu` is the constant (aka. λ)
  - Example: `t.test(IQ_values, mu=100)`
- Two samples:
  - The first two arguments are x_1 and x_2
  - `paired` controls if the samples are paired (`TRUE`) or independent (`FALSE`, default)
  - `var.equal` controls if variances are assumed to be equal
  - Example: `t.test(x1, x2, var.equal=TRUE)`

Effect size calculations

The library `effsize` provides the function `cohen.d()` to estimate the effect size of a $t-$test:

```
library(effsize)
cohen.d(IQ_values, NA, mu=100)  # no second sample -> NA
cohen.d(x1, x2)
```

If we want a shorter output with only the estimated effect size d, we can extract that information with the $ operator:

```
effect_size_result = cohen.d(x1, x2)
effect_size_result$estimate   # print only the estimate of d

cohen.d(x1, x2)$estimate   # short form
```

t-test (one sample)

We test whether the average depression score `depr` at T_1 differs from $\lambda = 19$:

$H_0 : \mu = 19, \quad H_1 : \mu \neq 19$

```
t.test(df_wide$MZP1, mu=19)
```

```
##
##  One Sample t-test
##
## data:  df_wide$MZP1
## t = 1.8436, df = 34, p-value = 0.07398
## alternative hypothesis: true mean is not equal to 19
## 95 percent confidence interval:
##  18.64619 26.26810
## sample estimates:
## mean of x
##  22.45714
```

```
cohen.d(df$MZP1, NA, mu=19)$estimate
```

```
## [1] 0.31162
```

t-test (paired variables)

We test whether the patient's depression scores `depr` *change* between T_1 and T_2:

$H_0 : \bar{x_d} = 0, \quad H_1 : \bar{x_d} \neq 0$

```
t.test(df_wide$MZP1, df_wide$MZP2, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  df_wide$MZP1 and df_wide$MZP2
## t = 3.5173, df = 34, p-value = 0.00126
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.592351 5.950506
## sample estimates:
## mean of the differences
##                3.771429
```

We test whether the patient's depression scores `depr` *decrease* between T_1 and T_2:

$H_0 : \bar{x_d} < 0, \quad H_1 : \bar{x_d} > 0$

```
t.test(df_wide$MZP2, df_wide$MZP1, paired=TRUE, alternative="less")
```

```
##
##  Paired t-test
##
## data:  df_wide$MZP2 and df_wide$MZP1
## t = -3.5173, df = 34, p-value = 0.0006298
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1.958332
## sample estimates:
## mean of the differences
##               -3.771429
```

t-test (two sample)

We test whether the means of x_1 and x_2 differ:

$H_0 : \bar{x_1} = \bar{x_2}, \quad H_1 : \bar{x_1} \neq \bar{x_2}$

```
x1 = rnorm(n=20, mean=98, sd=15)
x2 = rnorm(n=20, mean=100, sd=15)
t.test(x1, x2, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  x1 and x2
## t = 0.85752, df = 38, p-value = 0.3965
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.727119 16.614474
## sample estimates:
## mean of x mean of y
## 100.87880  95.93512
```

```
cohen.d(x1, x2)$estimate
```

```
## [1] 0.2711719
```