R Workshop: T-test

Brier Gallihugh, M.S.

2023-06-19

Table of contents

An Introduction to T-tests

Assumptions of T-tests

Normality of Residuals

```
library(tidyverse)

data <- starwars %>%
  filter(sex == "male" | sex == "female")                    ①

model <- lm(height ~ sex, data = data)                       ②

residuals <- data.frame(res = residuals(model))              ③

problem <- residuals %>% filter(res > 2.5 | res < -2.5)      ④
```

```
nrow(problem)/nrow(data)
```

① Filtering for male the female using the `filter()` function
② Running a linear regression (that is a t-test here) to get residuals
③ Calculate residuals for the observations
④ Find potentially problematic observations

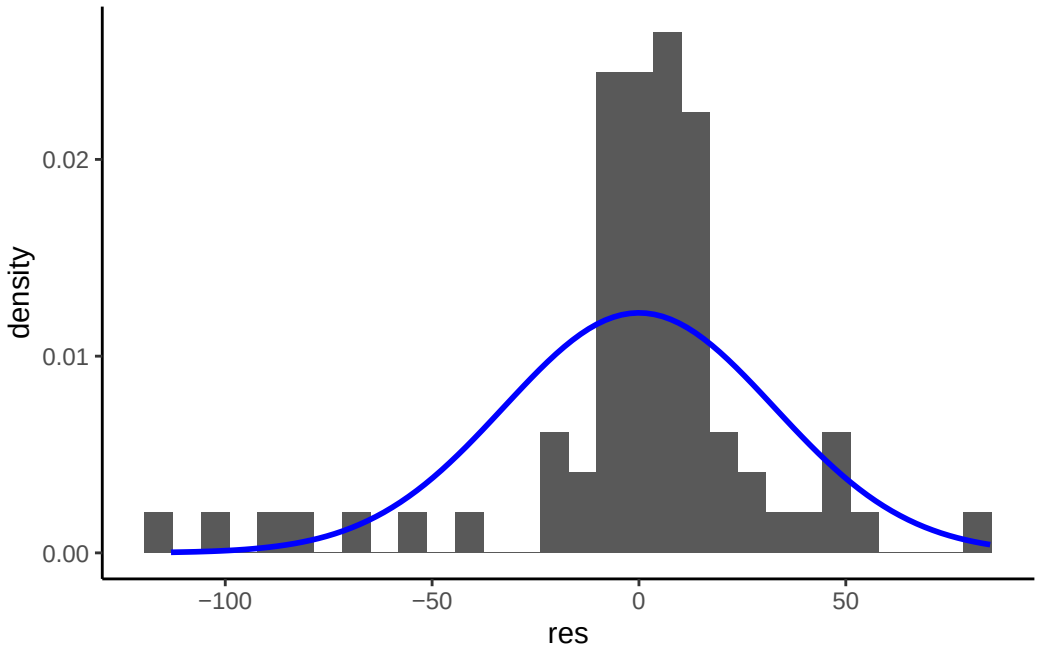
```
[1] 0.8157895
```

Graphical Depiction of Normality of Residuals

```
residual_graph <- ggplot(residuals,aes(x = res)) +                    ①
  geom_histogram(aes(y=after_stat(density))) +                        ②
  stat_function(fun = dnorm,                                          ③
                args = list(mean = mean(residuals$res),
                            sd = sd(residuals$res)),
                col = "blue",
                linewidth = 1) +
  theme_classic()

print(residual_graph)                                                 ④
```

① We are plotting the residuals here. We give ggplot a `geom` (i.e., histogram)
② We also give some other arguments like a density distribution.
③ Here we are basically providing what is needed to draw a normal distribution given the data using the `stat_function()` function. The `col` and `linewidth` arguments simply change the colr and size of the normal curve. The `theme_classic()` just changes some aesthetic things. I personally prefer this theme for all ggplot2 graphs
④ `print` will show us the graph output



Statistical Depiction of Normality of Residuals

We can also test the assumption statistically using the `shapiro.test()` function here

```
shapiro.test(residuals$res)
```

```
    Shapiro-Wilk normality test

data:  residuals$res
W = 0.84515, p-value = 3.515e-07
```

Homogeneity of Variance

Homogeneity of variance is important even for a basic t-test. Below is how we might go about testing this assumption.

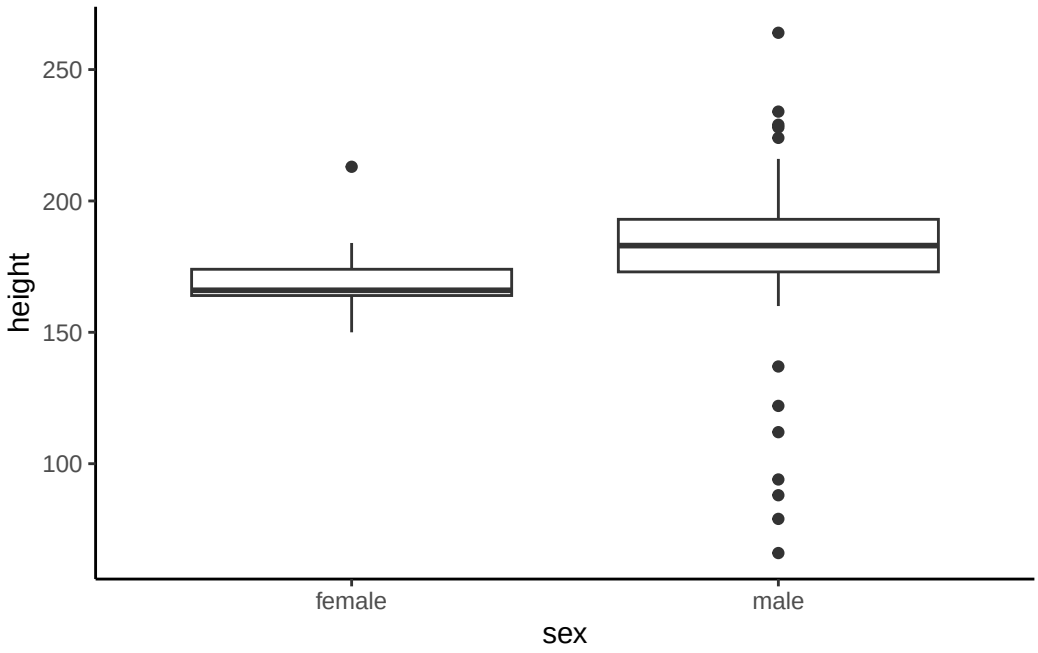
Graphical Depiction of Homogeneity of Variance

```
variance_boxplot <- ggplot(data,aes(x = sex,                    ①
                                    y = height)) +
```

```
  geom_boxplot() +                                                    ②
  theme_classic()

print(variance_boxplot)
```

① Graphically we can represent this as a boxplot with the group variable as the `x` and the outcome as the `y`. We see this here
② We again provide a `geom` for ggplot2 to use and provide a `theme()` choice here



Statistical Depiction of Homogeneity of Variance

We can also test the assumption using the Bartlett test. This can be shown below

```
bartlett.test(height ~ sex,data)
```

```
    Bartlett test of homogeneity of variances

data:  height by sex
Bartlett's K-squared = 11.126, df = 1, p-value = 0.0008511
```

Running a T-test

```r
t.test(height ~ sex, data = data)                                    ①
```

① The `t.test()` function will take a DV and IV argument as well as the dataframe used. We can see this here

```
    Welch Two Sample t-test

data:  height by sex
t = -1.5876, df = 55.148, p-value = 0.1181
alternative hypothesis: true difference in means between group female and group male is not e
95 percent confidence interval:
 -22.256861   2.579668
sample estimates:
mean in group female   mean in group male
            169.2667             179.1053
```