T-test, ANOVA, Mann-Whitney, Kruskal Wallis

Jason Hodgson

2024-02-28

# Necessary libraries

In this tutorial we will use: ggplot2. Run the following code to be sure you have the necessary dependencies installed.

# Check if ggplot2 is installed  
if (!requireNamespace("ggplot2", quietly = TRUE)) {  
 # If not installed, install ggplot2  
 install.packages("ggplot2")  
}  
  
  
# Load the installed packages  
library(ggplot2)

# Setting the working directory

This tutorial assumes that you have created a folder called “**week\_7\_t-test\_anova**” within your “**PfR\_stats\_portfolio**” folder, and have created “**data**” and “**analysis**” folders inside “**week\_7\_t-test\_anova**”. Save the class data file “**form-1\_\_anthropometrics.csv**” in the “**data**” folder, and save this .Rmd file in the “**analysis**” folder. Once you have set everything up, set the working directory to “**week\_7\_t-test\_anova**”, by choosing “**Session -> Set Working Directory -> Choose Directory**”.

Check that your working directory is set correctly using the getwd() function. You should see the full file path of your working directory.

# check your working directory  
getwd()

## [1] "/Users/jahodgson/Dropbox/Teaching/R\_resources/bioinformatics/R/statistics/4\_t-test\_ANOVA\_Mann-Whitney\_Kruskal-Wallis"

# Loading the data

# load the data from the .csv file  
d <- read.csv(file="../data/form-1\_\_anthropometrics.csv", header=TRUE, stringsAsFactors = TRUE)  
  
# check the structure of the data frame the str() function describes the full structure of a data frame, including the number of variables, and observations, and the class() of each variable  
str(d)

## 'data.frame': 81 obs. of 24 variables:  
## $ ec5\_uuid : Factor w/ 81 levels "01b70f78-552b-4bea-ad7b-f77ae55b01f7",..: 11 4 52 58 55 8 66 3 33 65 ...  
## $ created\_at : Factor w/ 81 levels "2024-02-05T11:40:46.000Z",..: 81 80 79 78 77 76 75 74 73 72 ...  
## $ uploaded\_at : Factor w/ 80 levels "2024-02-05T11:43:08.000Z",..: 80 79 77 76 74 73 72 71 70 69 ...  
## $ title : Factor w/ 61 levels "00:43 05/02/2024",..: 9 33 31 29 8 7 16 59 5 57 ...  
## $ X1\_Time : Factor w/ 59 levels "00:43","09:00",..: 9 31 29 27 8 7 14 57 5 55 ...  
## $ X2\_Date : Factor w/ 7 levels "03/09/2002","04/02/2024",..: 3 3 1 6 6 6 3 5 3 5 ...  
## $ X3\_Gender : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 1 1 1 1 ...  
## $ X4\_Age\_in\_years : int 21 19 21 19 20 20 22 18 20 21 ...  
## $ X5\_Height\_in\_cm : int 167 175 162 170 160 182 164 154 168 164 ...  
## $ X6\_Weight\_in\_kg : num 77 60 56 63 59 97.7 62.2 57 56 59 ...  
## $ X7\_Activity\_Level\_NAS: Factor w/ 7 levels "0 - Avoid walking or exertion, always use the elevator etc.",..: 4 6 2 2 6 7 5 5 4 6 ...  
## $ X8\_Smoker : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 1 1 1 ...  
## $ X9\_Time\_since\_last\_me: Factor w/ 7 levels ""," ﹥15 minutes",..: 3 3 5 3 4 3 7 6 1 3 ...  
## $ X10\_Time\_since\_caffei: Factor w/ 6 levels "","﹥ 15 mins",..: 4 3 3 3 3 3 2 3 1 2 ...  
## $ X11\_Time\_since\_exerci: int 30 90 1 24 1 24 2 9 NA 54 ...  
## $ X12\_Body\_temperature\_: num 39 34.6 36 36.6 37.5 37 36.3 37 36.1 36.9 ...  
## $ X13\_waist\_measurement: int 82 70 63 70 79 100 NA 73 NA 84 ...  
## $ X14\_hips\_measurement\_: int 101 90 104 93 88 114 NA 100 NA 109 ...  
## $ X15\_Heart\_rate\_bpm\_me: int 88 90 72 85 72 70 86 80 NA 57 ...  
## $ X16\_Systolic\_pressure: int NA 65 105 93 NA NA 120 100 125 101 ...  
## $ X17\_Diastolic\_pressur: int NA 120 50 66 NA NA 65 100 107 69 ...  
## $ X18\_Blood\_oxygen\_leve: int 94 90 94 NA NA NA 100 85 NA 98 ...  
## $ X19\_Heart\_rate\_bpm\_me: int 88 64 98 NA NA NA 72 110 NA 57 ...  
## $ X20\_Blood\_glucose\_mM : num 7.2 5 7.4 5.9 5.2 5.5 6.4 13 4.7 4.2 ...

# Functions used in this tutorial

getwd()

read.csv()

str()

shapiro.test()

t.test()

wilcox.test()

aov()

kruskal.test()

print()

ggplot2::ggplot()

ggplot2::geom\_histogram()

ggplot2::geom\_boxplot()

ggplot2::geom\_violin()

# Working with a numerical dependent variable, and a categorical independent variable

It is often the case that your experimental design will involve varying a treatment condition (e.g. *drug A* versus *placebo* controll), and measuring their effect on a measured numerical variable (e.g. *systolic blood pressure*). Alternatively, you might be interested in testing a hypothesis about the effect of some categorical variable (e.g. *sex*) on some measured variable (e.g. *height*) in data that you have collected. In this case you can use a T-test, ANOVA, Mann-Whitney, or Kruskal-Wallis test. The conditions for which of these you should choose are as follows:

* **t-test**: normally distributed dependent variable, two categories in the grouping variable
* **Mann-Whitney**: non-normally distributed dependent variable, two categories in the grouping variable
* **ANOVA**: normally distributed dependent variable, three or more categories in the grouping variable
* **Kruskal-Wallis**: non-normally distributed dependent variable, three or more categories in the grouping variable

# Checking if your data are normal

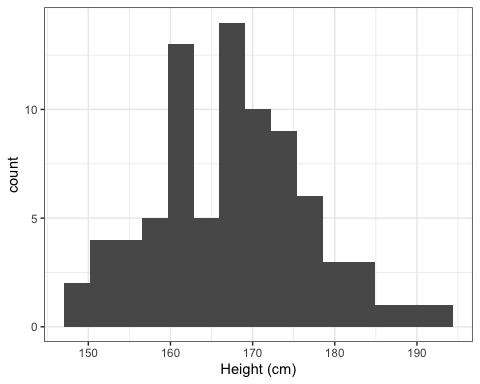
The first step in formally analysing your data are evaluating whether your numerical variable is how your numerical variable is distributed. If your data are normal, then it is best to use one of the parametric tests (T-test, ANOVA) because these have greater sensitivity (greater statistical power). This means you are less likely to fail to reject the null hypothesis even when it is false (Type 2 error).

## Visually examining your data

Begin by visualising the distribution of your numerical variable with a histogram, to check if your data are bell-shaped, or whether there is obvious skew. We will use the ggplot2 R package for producing publication quality histogram plots.

### Histogram of height

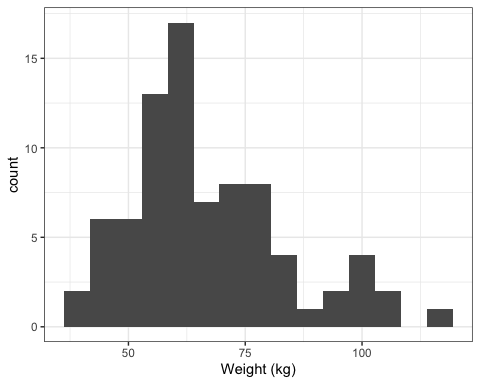
# choose a binwidth for your plot. The binwidth is the with of the bars in histogram, and is in the same scale as your variable (e.g. height in cm, then binwidth is in cm). The choice of binwidth is up to you, and you might want to experiment a bit to find a binwidth that displays the data best. As a rule of thumb, if the sample size is small you will want a bigger binwidth, but your binwidth should never be smaller than the precision of your measurment. For example if you measured height with a precision of 1 cm, having a binwidth of 0.1 makes no sense because the bars cannot possibly be touching since you will have bins for values that cannot be present in the data.  
  
# A sensible binwidth for small sample size data (n < 100) is often a third of the standard deviation. For very large data sets, set the binwidth equal to the precision of the measurment.  
  
height\_binwidth <- sd(d$X5\_Height\_in\_cm, na.rm=TRUE)/3  
  
# create a histogram of a the height variable  
P\_height <- ggplot(data=d, aes(x=X5\_Height\_in\_cm)) + # define the variables in the plot  
 geom\_histogram(binwidth=height\_binwidth) + # call the histogram function and set the binwidt  
 theme\_bw() + # choose a cleaner plot theme  
 xlab("Height (cm)") # relable the x axis  
  
# print the plot to the document  
print(P\_height)



In this plot the data resemble a normal distribution for a relatively small sample (n = 81).

### Histogram of weight

# define a binwidth  
weight\_binwidth <- sd(d$X6\_Weight\_in\_kg, na.rm=TRUE)/3  
  
# make the plot  
P\_weight <- ggplot(data=d, aes(x=X6\_Weight\_in\_kg)) + # define the variables in the plot  
 geom\_histogram(binwidth=weight\_binwidth) + # call the histogram function and set the binwidth  
 theme\_bw() + # choose a cleaner plot theme  
 xlab("Weight (kg)") # relable the x axis  
  
# print the plot to the document  
print(P\_weight)



In this case the data look like they may be left-skewed. In other words there are more small values in the data, and a long tail of large values.

## Formally testing for normality: The Shapiro-Wilk test

There are two commonly used formal tests of normality: the Shapiro-Wilk test, and the Kolmogorov-Smirnov tests. These both test the hypothesis that the sample data are drawn from a normal distribution. This makes the alternative hypothesis that the data are not normal. Thus, the interpretation of the computed p-value is as follows:

* p-value > 0.05: data are normal
* p-value < 0.05: data are not normal

We will use the Shapiro-Wilk test to check for normality. This is performed using the shapiro.test() function.

### Testing for normality of height

# test for normality of height  
height\_normality\_test <- shapiro.test(d$X5\_Height\_in\_cm)  
  
# print the results to the document  
print(height\_normality\_test)

##   
## Shapiro-Wilk normality test  
##   
## data: d$X5\_Height\_in\_cm  
## W = 0.98275, p-value = 0.3473

In this test, the p-value is greater than 0.05, and we accept the null hypothesis that the height data are normally distributed.

### Testing for normality of weight

# test for normality of weight  
weight\_normality\_test <- shapiro.test(d$X6\_Weight\_in\_kg)  
  
# print the results to the document  
print(weight\_normality\_test)

##   
## Shapiro-Wilk normality test  
##   
## data: d$X6\_Weight\_in\_kg  
## W = 0.93569, p-value = 0.0005227

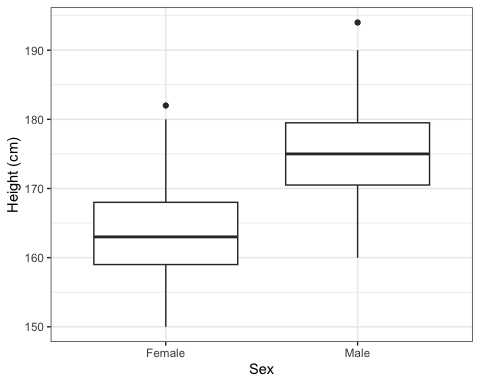
In this test, the p-value is less than 0.05 and we reject the null hypothesis that the weight data are normally distributed. In this case, we cannot treat the weight data as normal, and cannot use parametric statistical tests that assume normality.

# Plotting a categorical variable versus a numerical variable

## Box plots

### Traditional box plot

# create the plot  
P\_box\_1 <- ggplot(data=d, aes(x=X3\_Gender, y=X5\_Height\_in\_cm)) +  
 geom\_boxplot() +  
 theme\_bw() +  
 xlab("Sex") +  
 ylab("Height (cm)")  
  
# print the plot to the document  
print(P\_box\_1)



### Notched boxplot

A notched boxplot is similar to traditional boxplot, however notches are added to the boxes. The notches can be used to compare the groups. As a general rule, if the notches do not overlap, it suggests that the medians between the two groups are significantly different. In this case, it is likely that a formal hypothesis test of the difference between means will be significant.

# create the plot  
P\_box\_2 <- ggplot(data=d, aes(x=X3\_Gender, y=X5\_Height\_in\_cm)) +  
 geom\_boxplot(notch=TRUE) +  
 theme\_bw() +  
 xlab("Sex") +  
 ylab("Height (cm)")  
  
# print the plot to the document  
print(P\_box\_2)

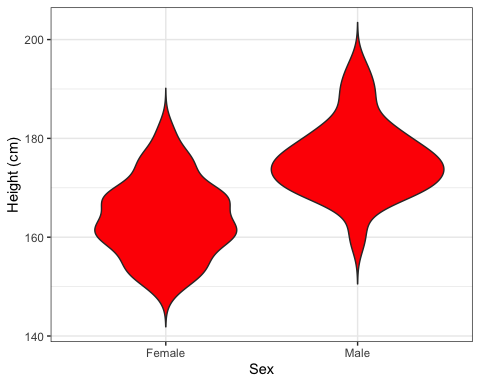


In this plot we see no overlap between the notches for the groups, and we would expect there to be a significant difference in height between males and females in our data if tested formally.

## Violin plots

A violin plot gives a very good representation of the shape of the distribution of the underlying data. The width of the “violin” indicates the frequency of the observations. Normal data should look like a bell curve turned on its side, and mirrored.

# create the plot  
P\_violin <- ggplot(data=d, aes(x=X3\_Gender, y=X5\_Height\_in\_cm)) +  
 geom\_violin(trim=FALSE, fill="red") + # setting trim to FALSE displays the full range  
 theme\_bw() +  
 xlab("Sex") +  
 ylab("Height (cm)")  
  
# print the plot to the document  
print(P\_violin)



In this plot we can see that the most frequent heights for females and males do not line up, suggesting a difference in the means between these groups.

# Performing a t-test

Performing any version of the t-test in R is easily done with the t.test() function. We will illustrate one and two-tailed versions of the unpaired t-test by testing for sex differences in average height. In our data, we have already seen that height is normally distributed, and thus does not violate the assumptions of the t-test.

## Two-tailed unpaired t-test

The two-tailed version of the hypothesis is that there is a difference in average heights between males and females. For the two-tailed version of the test, it does not matter whether males or females are taller on average, it only matters whether they are different.

# calculate the t-test of the effect of sex on height  
two\_tailed\_height <- t.test(d$X5\_Height\_in\_cm ~ d$X3\_Gender, alternative="two.sided")  
  
# report the test results to the document  
print(two\_tailed\_height)

##   
## Welch Two Sample t-test  
##   
## data: d$X5\_Height\_in\_cm by d$X3\_Gender  
## t = -6.6674, df = 50.842, p-value = 1.85e-08  
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0  
## 95 percent confidence interval:  
## -15.626286 -8.393294  
## sample estimates:  
## mean in group Female mean in group Male   
## 163.8364 175.8462

In this test the males are reported to have an mean height of 175.8 cm, and females a mean height of 163.8 cm, and this difference was found to be significant (t = -6.67, df=50.8, p=1.85e-08). Note when reporting a t-test, you must give the t value, the degrees of freedom, and the p value.

## One-tailed unpaired t-test

If you have an *a priori* reason to believe either males or females are taller, you should can use a one-tailed version of the t-test. In this case we will test the hypothesis that males are taller than females on average. To perform a one-tailed test we have to specify either "greater" or "less" to the alternative argument of the t.test() function. The "less" and "greater" indicators refer to the order of the factor levels in the grouping variable. If "Female" is the first level and "Male" is the second level then we want to specify "less". On the other hand if "Male" is the first level and "Female" is the second level then we want to specify "greater". You can check the order of the factor levels with the levels() function: enter levels(d$X3\_Gender) on the console. Alternatively, you can look at the output of the str() function when we described the data frame at the beginning of the tutorial. It happens to be the case that "Female" is the first level, so we want to indicate "less".

# calculate the one-tailed t-test of the effect of sex on height  
one\_tailed\_height <- t.test(d$X5\_Height\_in\_cm ~ d$X3\_Gender, alternative="less")  
  
# print the test result to the document  
print(one\_tailed\_height)

##   
## Welch Two Sample t-test  
##   
## data: d$X5\_Height\_in\_cm by d$X3\_Gender  
## t = -6.6674, df = 50.842, p-value = 9.251e-09  
## alternative hypothesis: true difference in means between group Female and group Male is less than 0  
## 95 percent confidence interval:  
## -Inf -8.991957  
## sample estimates:  
## mean in group Female mean in group Male   
## 163.8364 175.8462

The one-tailed version of the test shows that males significantly taller than females in the class data (t = -6.67, df = 50.84, p = 9.25e-09). Notice the t value, and the degrees of freedom are the same as the two-tailed test, but the p-value is less. This is because the significant area of the tail is greater in the one-tailed hypothesis test, making the t-value more extreme.

# Performing a Mann-Whitney U test

The Mann-Whitney U test is used when you want to perform a t-test, however your data are not normally distributed. We will test hypotheses about a sex difference in body weight. This is performed in exactly the same way as the t-test, however it uses the wilcox.test() function instead. Another name for the Mann-Whitney U test is the Wilcoxon Rank Sums test.

## Two-tailed Mann-Whitney U test

The two-tailed hypothesis is that there is a difference in body weight between males and females. The two tailed test is agnostic about which sex is heavier, it only matters that there is a difference.

# calculate the Mann-Whitney U test  
two\_tailed\_weight <- wilcox.test(d$X6\_Weight\_in\_kg ~ d$X3\_Gender, alternative="two.sided")  
  
# print the results to the document  
print(two\_tailed\_weight)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: d$X6\_Weight\_in\_kg by d$X3\_Gender  
## W = 262.5, p-value = 4.8e-06  
## alternative hypothesis: true location shift is not equal to 0

The Mann-Whitney U test shows that there is a significant difference in weight between males and females (W = 262.5, p = 4.8e-06). Note that when reporting the test results you should report the W test statistic as well as the p-value.

## One-tailed Mann-Whitney U test

If you have an *a priori* reason to believe one of the groups is greater than the other, you can specify a one-tailed hypothesis. We will test the hypothesis that males are heavier than females in the class data. Similar to the t.test() function, you must specify either "less" or "greater" according to the order of the factor levels. In our case "Female" is the first factor level so we will specify "less". See the description of the one-tailed t-test if this is confusing.

# calculate the Mann-Whitney U test  
one\_tailed\_weight <- wilcox.test(d$X6\_Weight\_in\_kg ~ d$X3\_Gender, alternative="less")  
  
# print the results to the document  
print(one\_tailed\_weight)

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: d$X6\_Weight\_in\_kg by d$X3\_Gender  
## W = 262.5, p-value = 2.4e-06  
## alternative hypothesis: true location shift is less than 0

# Performing an ANOVA

An ANOVA is the appropriate test to check for a difference in means if the data are normal and if there are more than three groups in the grouping variable. The aov() function is used to calculate an ANOVA. This function fits the ANOVA model. To summarise the results including seeing the p-value you need to use the summary() function on the fitted model. We previously saw that the data for height is normal, and thus fits the assumptions of the ANOVA test. We will test the (rather nonsensical) hypothesis that activity level effects height. There are seven levels to the activity level variable.

# use the aov() function to fit the model  
my\_anova <- aov(X5\_Height\_in\_cm ~ X7\_Activity\_Level\_NAS, data=d)  
  
# summarise the model  
summary(my\_anova)

## Df Sum Sq Mean Sq F value Pr(>F)  
## X7\_Activity\_Level\_NAS 6 835 139.09 1.618 0.154  
## Residuals 74 6361 85.96

In this test, no significant relationship was found between activity level and height in the class (F = 1.618, df = 6, p = 0.154). Note that when reporting the results of an ANOVA you must report the F statistic, the degrees of freedom, and the p-value.

# Performing a Kruskal-Wallis test

The non-parametric version of the ANOVA is the Kruskal-Wallis test. This test can be performed on non-normal data broken into more than three categories by a grouping variable. The test is performed with the kruskal.test() function. We will test the hypothesis that there is a difference in body weight among people with different activity levels.

# calculate the kruskal-wallis test  
my\_kruskal <- kruskal.test(d$X6\_Weight\_in\_kg ~ d$X7\_Activity\_Level\_NAS)  
  
# print the results to the document  
print(my\_kruskal)

##   
## Kruskal-Wallis rank sum test  
##   
## data: d$X6\_Weight\_in\_kg by d$X7\_Activity\_Level\_NAS  
## Kruskal-Wallis chi-squared = 9.31, df = 6, p-value = 0.1569

The results show that there is no evidence for a difference in body weight associated with differences in activity level in the class data (Kruskal-Wallis chi-squared=9.31, df = 6, p = 0.157). Note that when reporting results of this test you need to give the test statistic, the degrees of freedom, and the p-value.

# Class Exercises

## Exercise 1. Test the hypothesis that sex affects body temperature.

### Step 1. Plot and interpret the histogram of body temperature

*Insert R code and interpretation here*

### Step 2. Test for normality of body temperature using the Shapiro-Wilk test

*Insert R code and interpretation here*

### Step 3. Visualise the relationship between sex and body temperature with a boxplot or violin plot

*Insert R code and interpretation here*

### Step 4. Formally test and interpret the relationship between sex and body temperature using either a t-test, or Mann-Whitney U test depending upon whether body temperature is normal or not.

*Insert R code and interpretation here*

## Exercise 2: Test the hypothesis that activity level affects heart rate.

### Step 1. Plot and interpret the histogram of heart rate

*Insert R code and interpretation here*

### Step 2. Test for normality of heart rate using the Shapiro-Wilk test

*Insert R code and interpretation here*

### Step 3. Visualise the relationship between activity level and heart rate with a boxplot or violin plot

*Insert R code and interpretation here*

### Step 4. Formally test and interpret the relationship between heart rate and activity level using either a ANOVA, or Kruskal-Wallis test depending upon whether heart rate is normal or not.

*Insert R code and interpretation here*