A BEGINNER
GUIDE TO

t-test and ANOVA (Analysis of Variance) in R Programming

Made by Habbee

Overview

T-test:

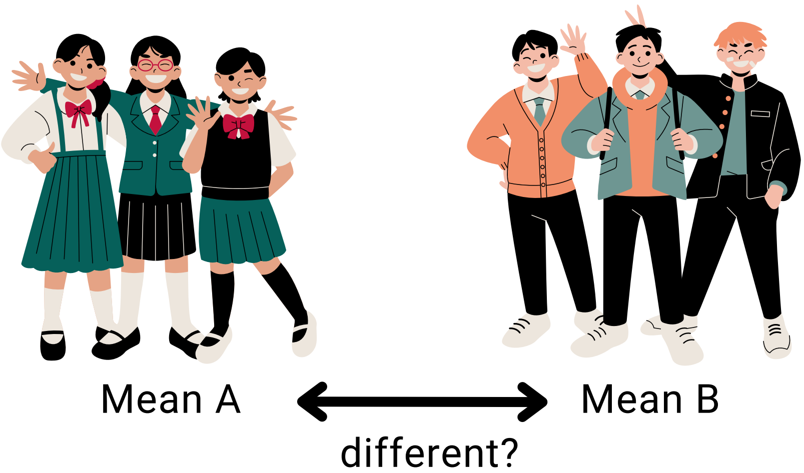
- Independent t-test.
- Paired t-test.

F-test:

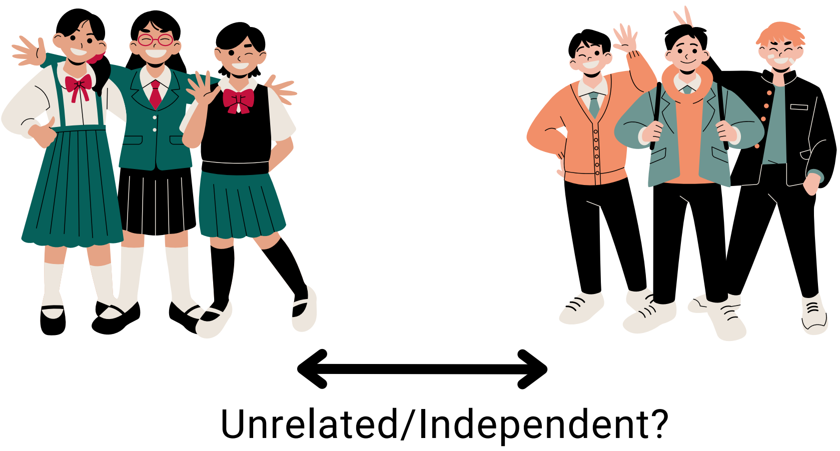
- One-way Analysis of Variance (ANOVA).
- Two-way Analysis of Variance (ANOVA).

1. t-test:

To test **difference in means** for two *small* samples (n < 30) from populations that are approximately *normal*.
(The two small samples are representatives of their parent populations).



To test the **linear dependence** to check if the two *small* samples are unrelated/independent.



1. t-test:

1.1. Independent samples t-test

is applied when we want to test **differences between the means/averages** of two completely **independent** groups (one does not affect the other).

For instance, Ty goes on a three-mile run with his kids every morning. He wanted to test if his son's running time (in minutes) is *significantly lower* than his daughter's — meaning the boy can run faster. To test the theory, he recorded their running times everyday for a week as given in the following table:

Son's running time (in minutes)	Daughter's running time (in minutes)
20	30
22	26
16	24
21	19
15	17
17	19
16	21

Running time records (in minutes).

First step, create the running time records in Rstudio.

```r
#Independent t-test
Kids <- c(rep(c("Son","Daughter"), each=7))
Minutes <- c(20,22, 16, 21, 15, 17, 16, 30, 26, 24, 19, 17, 19, 21)
RunData <-data.frame(Kids,Minutes)
RunData
```

Kids <chr>	Minutes <dbl>
Son	20
Son	22
Son	16
Son	21
Son	15
Son	17
Son	16
Daughter	30
Daughter	26
Daughter	24

1-10 of 14 rows Previous 1 2 Next

Import the data into R.

We name the independent variables as “Son” and “Daughter”. Since R reads data *alphabetically* , the daughter’s data is always processed before son’s, as the letter D goes before S in the alphabet; Thus, our updated **alternative hypothesis** *Ha* now has become *μ (daughter) > μ (son),* which is still equivalent to Ty’s theory — “his son’s running time is significantly lower than his daughter’s” .

H0: μ (daughter) = μ (son)

Ha: μ (daughter) > μ (son)

```r
t.test(Minutes~Kids,data = RunData, alternative = "greater")

#Alternative includes "two.sided", "greater", "less"
```

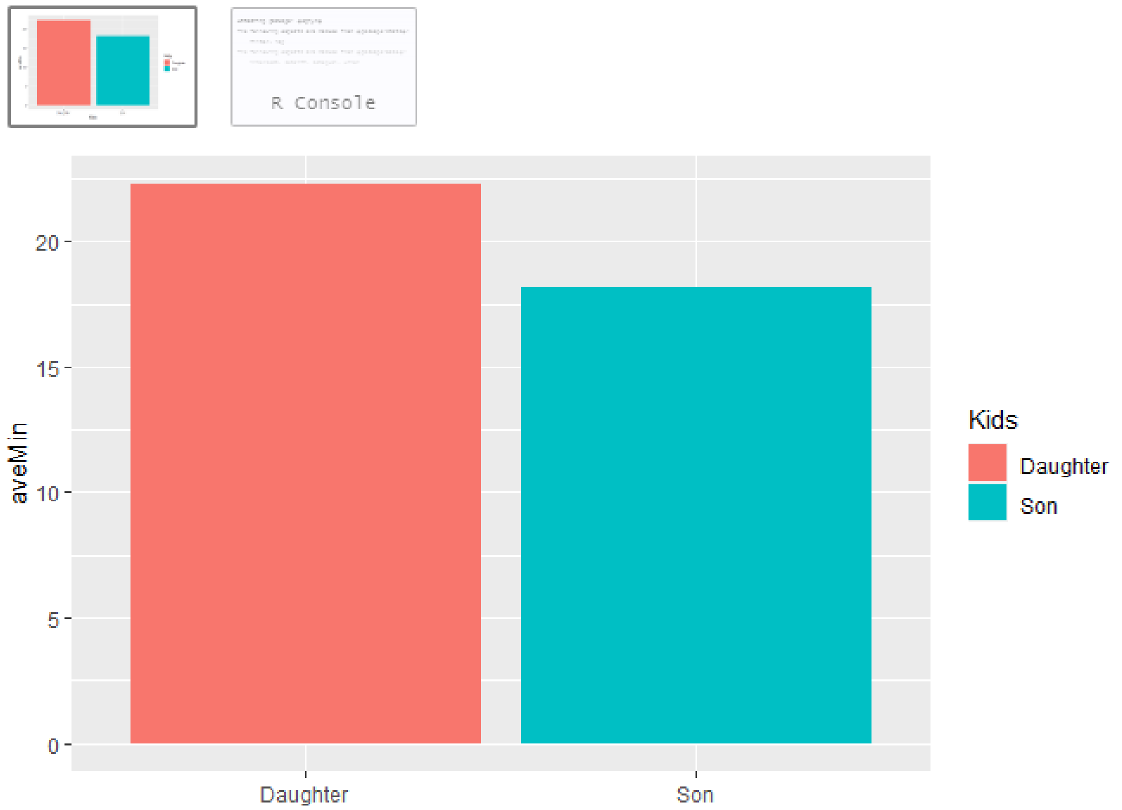
```
	Welch Two Sample t-test

data:  Minutes by Kids
t = 2.0337, df = 9.887, p-value = 0.03485
alternative hypothesis: true difference in means between group Daughter and group
Son is greater than 0
95 percent confidence interval:
 0.4464573       Inf
sample estimates:
mean in group Daughter      mean in group Son 
              22.28571               18.14286 
```

Independent t-test syntax.

From the result, **t-statistic** is 2.0337, and **p-value** = 0.03485, meaning it is *less* than 0.05 (using the 0.05 significance level) ; therefore, *H0* is rejected. There is enough sufficient evidence to support *Ha* that the daughter has a higher mean running time than the son.

In addition, R also calculates both **the means of the daughter's running time** (22.29 minutes) and **son's** (18.14 minutes); hence, we can conclude that Ty's son is faster when he runs the three-mile route! Let's view it in visualization!

```
library(ggplot2)
library(dplyr)
RunData%>%
  group_by(Kids)%>%
  summarise(aveMin= mean(Minutes))%>%
  ggplot(aes(x = Kids,y = aveMin)) + geom_bar(stat="identity",aes(fill=Kids))
```



Last but not least, **sample sizes** for the two groups *sometimes* are not equally the same. For example, what if Ty's daughter got busy one morning and could not join the morning run with her brother and father during the week? The *sample size* for her running data would be 6 instead of 7!

If groups sizes differ ***greatly*** (**Homogeneity of Variance** is violated), that can cause the null hypothesis to be falsely rejected (**type I error**: reject *H0* when it is in fact true!)

1) t-test

1.2. Paired t-test:

is applied when we have **two dependent (paired)** samples from just one population and want to see if they are **significantly different** - useful for “*before and after*” situation.

Mean A (before) ⟷ Mean A (after)

different?

Example, Ty wants to test the difference in means of his kids’ heart rates *before* and *after* the three-mile run.

	Heart rate (in bpm)	
	Before	After
Son	72	90
Daughter	81	96

Heart rate records of “before” and “after” running 3 miles (in bpm).

Import the dataset into R for our paired t-test analysis.

~~~
#Paired t-test
at <-c(rep(c("Before","After"), each=2))
bpm <-c(72, 81, 90, 96)
heartRate <-data.frame(at, bpm)
heartRate
```
~~~

Description: df [4 x 2]

at <chr>	bpm <dbl>
Before	72
Before	81
After	90
After	96

4 rows

Import data into R.

H0: μ (before) = μ (after)

Ha: μ (before) ≠ μ (after)

~~~
t.test(bpm~at, data=heartRate, alternative="two.sided",paired=TRUE)
```
~~~

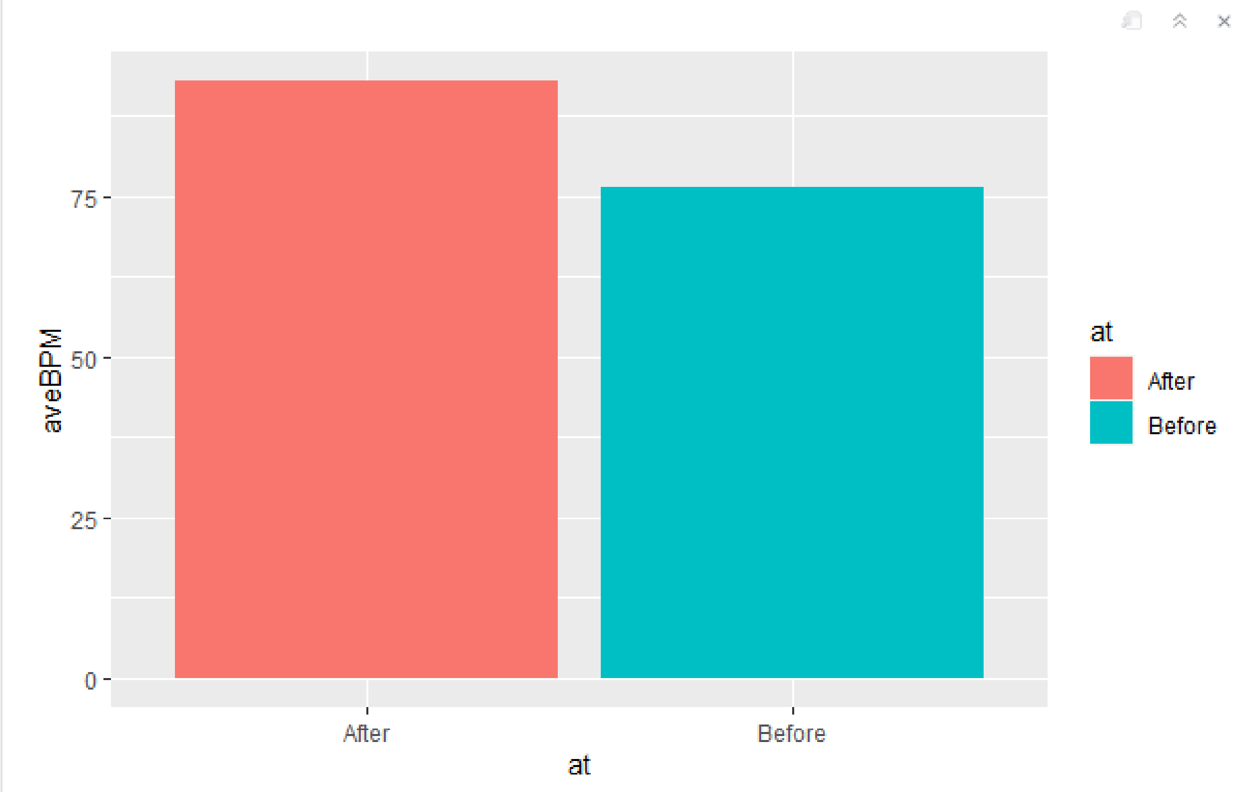
```
        Paired t-test

data:  bpm by at
t = 11, df = 1, p-value = 0.05772
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.559307 35.559307
sample estimates:
mean of the differences 
                   16.5 
```

Paired t-test syntax.

With **p-value** = 0.05772 (that is greater than 0.05), we fail to reject *H0* as we do *not* have enough sufficient evidence to support Ty's kids heart rates differ *significantly* (*statistically*) before and after the 3-mile run.

However, the result also shows that the **mean of the differences** is 16.5 bpm, and if we visualize our paired t-test, we can see the mean bpm from "*after*" running is higher than "*before*". Our hearts tend to beat faster per minute after we exercise!

```
heartRate%>%
  group_by(at)%>%
  summarise(aveBPM= mean(bpm)) %>%
  ggplot(aes(x = at,y = aveBPM)) + geom_bar(stat="identity",aes(fill=at))

```



Paired t-test visualization.

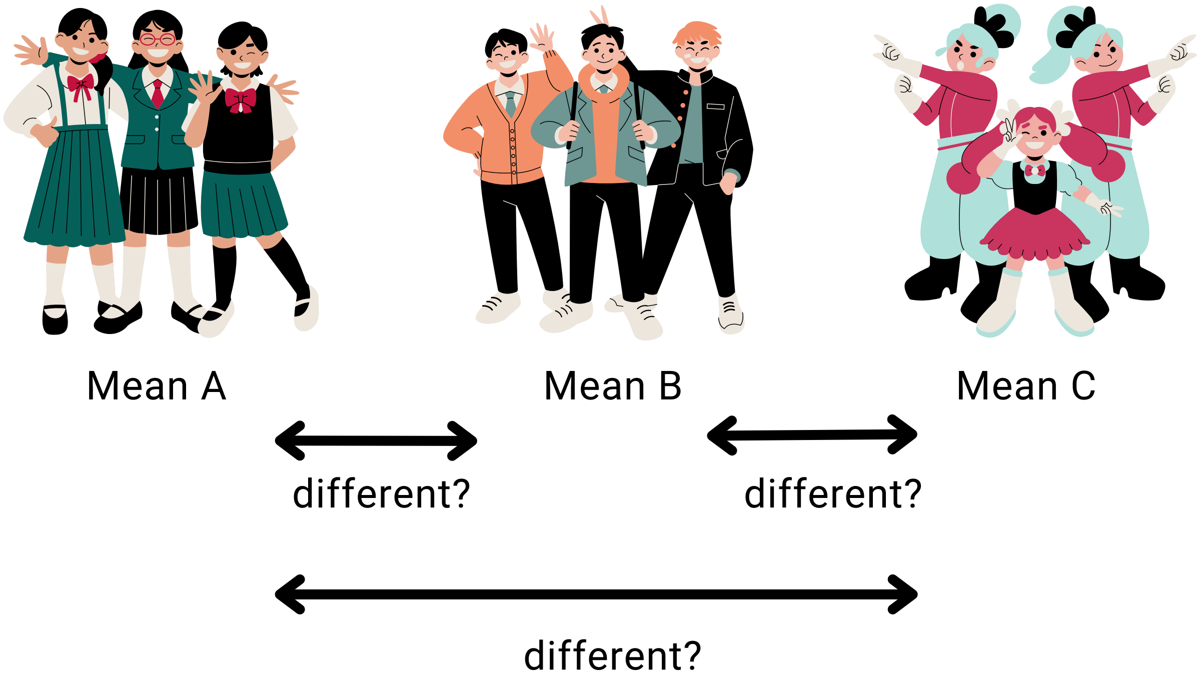
As a final point, **sample sizes** for the two measurements in *paired t-test* are *always* **identical** (equal variances), unlike *independent t-test*.

2. F-test:

Analysis of Variance (ANOVA)

works exactly like t-test but with more than two groups.

H0: μ (1) = μ (2) = μ (3)= … = μ (n)
Ha: at least two means are different.



Assumptions of ANOVA: Each groups of samples are *normally distributed* , have *equal variances*, and are *independent*.

2. F-test:

2.1. One-way ANOVA

is used to analyze the **difference between the means** of more than two groups.

Assume the **Dependent variable** (DV) is how many miles that a car can travel per gallon of fuel (mpg), and the **Independent variable** (IV) is different brands of cars . Apply an analysis of variance to test if the means are significantly different between them .

Toyota 4Runner	Subaru Crosstrek	Lexus RX350
19	28	20
17	30	23
16	32	25
20	33	24
17	31	21
19	27	22
15	29	24
21	30	21

Mpg records of Toyota 4Runner, Subaru Crosstrek, and Lexus RX350.

Let's let R read our mpg data.

```
#One way ANOVA

car <- c(rep(c("Toyota", "Subaru", "Lexus"), each=8))
mpg <- c(19, 17, 16, 20, 17, 19, 15, 21,
         28, 30, 32, 33, 31, 27, 29, 30,
         20, 23, 25, 24, 21, 22, 24, 21)
mpgData <-data.frame(car, mpg)
mpgData
```

Description: df [24 x 2]

car <chr>	mpg <dbl>
Toyota	19
Toyota	17
Toyota	16
Toyota	20
Toyota	17
Toyota	19
Toyota	15
Toyota	21
Subaru	28
Subaru	30

1-10 of 24 rows Previous 1 2 3 Next

Import data into R.

H0: μ (Toyota) = μ (Subaru) = μ (Lexus)

Ha: at least two means are different.

```
model <- aov(mpg~car, data= mpgData)
summary(model)
TukeyHSD(model)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)    
car          2    588  294.00   77.17 2.1e-10 ***
Residuals   21     80    3.81                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg ~ car, data = mpgData)

$car
               diff        lwr       upr     p adj
Subaru-Lexus    7.5   5.040175  9.959825 0.0000005
Toyota-Lexus   -4.5  -6.959825 -2.040175 0.0004259
Toyota-Subaru -12.0 -14.459825 -9.540175 0.0000000
```

One-way ANOVA syntax.

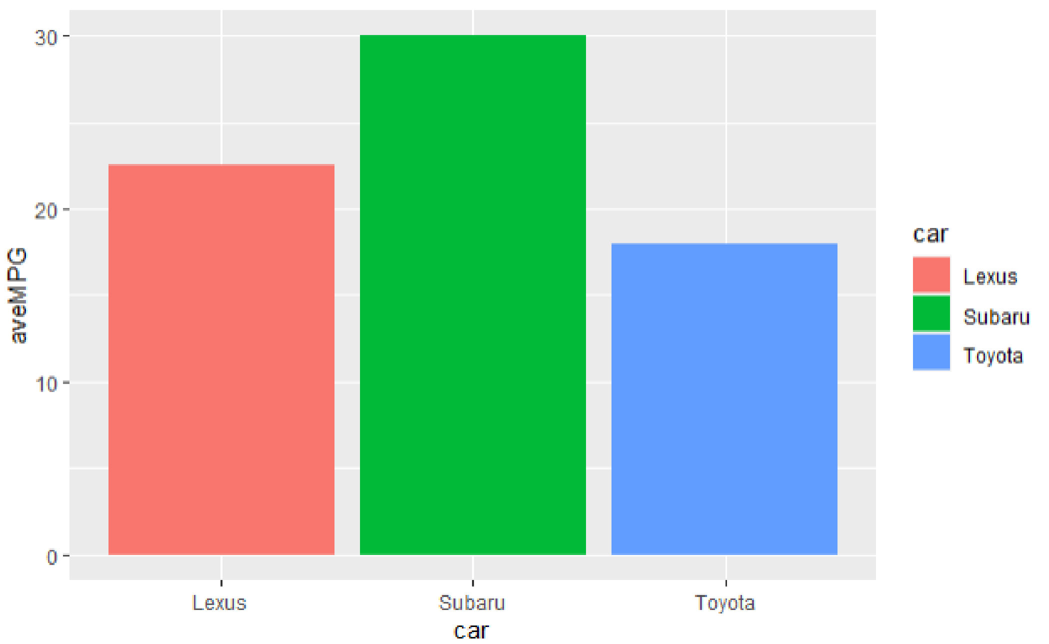
With our **F-statistic** is 77.17 and **p-value** is less than 0.05 (= 2.1e-10), we reject null hypothesis, and there is enough evidence to claim that at least two means are different.

.... but you may ask **which means are different**? The “*TukeyHSD(model)*” syntax helps us clarify that. Since the “**p-adj**” values between each pair of cars are < 0.05, we can state that there is a *significant difference* in average of mpg between Subaru and Lexus, Toyota and Lexus, and Toyota and Subaru, with Toyota 4Runner and Subaru differ the most in terms of mpg (“diff”= 12.0).

```
$car
                diff        lwr       upr     p adj
Subaru-Lexus     7.5   5.040175  9.959825 0.0000005
Toyota-Lexus    -4.5  -6.959825 -2.040175 0.0004259
Toyota-Subaru  -12.0 -14.459825 -9.540175 0.0000000
```

Means comparison.

```
mpgData%>%
  group_by(car)%>%
  summarise(aveMPG= mean(mpg)) %>%
  ggplot(aes(x = car,y = aveMPG)) + geom_bar(stat="identity",aes(fill=car))
```



Our one-way ANOVA visualization.

Last but not least, if the **confidence interval does *not* contain value 0** then there is a *significant difference* between two variables’ averages.

For example, the lower bound (lwr) and upper bound (upr) of Subaru-Lexus’ confidence interval are (5.0402, 9.9598), which do not consist of 0.

2. F-test:

2.2. Two-way ANOVA:

is applied when we want to analyze how two **Independent variables** (IV), in combination, ***affect*** a **Dependent variable** (DV) because we want to study if there is **an interaction** between the two IVs on our DV.

For instance, we want to know if the cars' mpg values mentioned above will differ when driven on highway and in the city.

The IVs now are car brands (Toyota, Subaru, and Lexus) and where they are being driven (in the city or on the highway), with our DV is mpg values.

	Toyota 4Runner	Subaru Crosstrek	Lexus RX350
City	14	28	20
	12	26	21
	15	26	19
Highway	19	33	27
	18	32	26
	19	33	24

The two-way ANOVA mpg dataset.

Here is how to create a two-way ANOVA data frame in R.

```
library(tidyverse)
library(ggplot2)

where <-c(rep(c("city", "highway"), each = 9))
brand <-c(rep(c("Toyota","Subaru","Lexus"),each =3))
mpg2 <- c(14, 12, 15,
          28, 26, 26,
          20, 21, 19,
          19, 18, 19,
          33, 32, 33,
          27, 26, 24)
Twoway <- data.frame(where, brand, mpg2)
Twoway
```

Description: df [18 x 3]

where <chr>	brand <chr>	mpg2 <dbl>
city	Toyota	14
city	Toyota	12
city	Toyota	15
city	Subaru	28
city	Subaru	26
city	Subaru	26
city	Lexus	20
city	Lexus	21
city	Lexus	19
highway	Toyota	19

1-10 of 18 rows   Previous 1 2 Next

Our two-way ANOVA mpg dataset in R.

We now have three different hypotheses to test, with the first one is:

***H0*: μ (Toyota) = μ (Subaru) = μ (Lexus)**
***Ha*: at least two means are different.**

```
model1 <-aov(mpg2~where + brand + where*brand, data=Twoway)
summary(model1)
```

```
            Df Sum Sq Mean Sq F value   Pr(>F)    
where        1  138.9  138.89 108.696 2.28e-07 ***
brand        2  546.8  273.39 213.957 4.12e-10 ***
where:brand  2    0.8    0.39   0.304    0.743    
Residuals   12   15.3    1.28                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-way ANOVA test syntax.

P-value of "*brand*" is 4.12e-10; we can claim that there is a *significant difference* of effect between driving the Toyota 4Runner, Subaru Crosstrek, and Lexus RX350 in terms of mpg, at least for two of the brands.

Next, our second hypothesis is:

H0: μ (city) = μ (highway)
Ha: μ (city) ≠ μ (highway)

Similar to our variable "*brand*", "*where*" we drive our cars is another factor that does have a *significant effect* on the mean difference of our miles per gallon because the **p-value** is less than 0.05 (= 2.28e-07).

In fact, we obtain *higher* mpg on highways than in the cities for majority of cars out there in the market.

Last but not least, our last hypothesis is:

H0: there is no interaction between what brand of car you drive and where you drive it.
Ha: there is an interaction between what brand of car you drive and where you drive it.

Our **test statistic value** is 0.0304 and **p-value** is 0.743. We fail to reject the null hypothesis, and there is not enough evidence to support the claim that there is *an interaction* between the cars brands and where you drive your car.

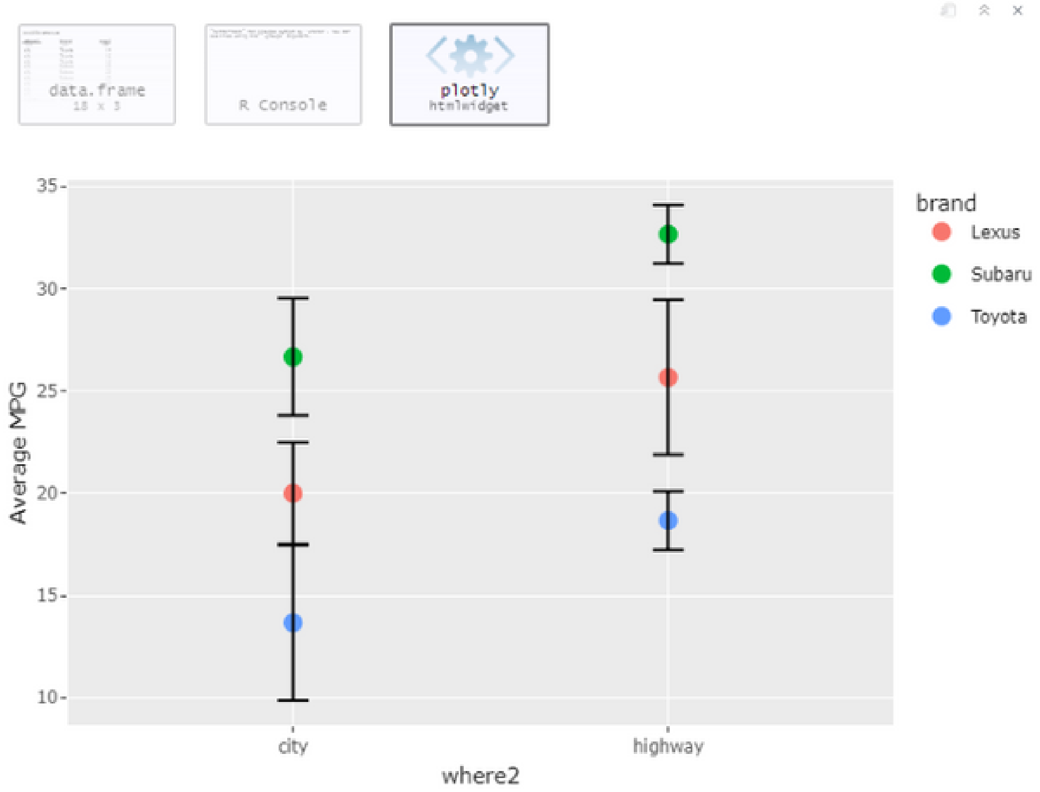
```
TukeyHSD(model1)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mpg2 ~ where + brand + where * brand, data = Twoway)

$where
                 diff      lwr     upr p adj
highway-city 5.555556 4.394531 6.71658 2e-07

$brand
                    diff        lwr        upr p adj
Subaru-Lexus    6.833333   5.092205   8.574461 6e-07
Toyota-Lexus   -6.666667  -8.407795  -4.925539 8e-07
Toyota-Subaru -13.500000 -15.241128 -11.758872 0e+00

$`where:brand`
                                diff        lwr        upr     p adj
highway:Lexus-city:Lexus        5.666667   2.566523   8.766810 0.0005535
city:Subaru-city:Lexus          6.666667   3.566523   9.766810 0.0001194
highway:Subaru-city:Lexus      12.666667   9.566523  15.766810 0.0000001
city:Toyota-city:Lexus         -6.333333  -9.433477  -3.233190 0.0001960
highway:Toyota-city:Lexus      -1.333333  -4.433477   1.766810 0.7020060
city:Subaru-highway:Lexus       1.000000  -2.100144   4.100144 0.8788715
highway:Subaru-highway:Lexus    7.000000   3.899856  10.100144 0.0000738
city:Toyota-highway:Lexus     -12.000000 -15.100144  -8.899856 0.0000002
highway:Toyota-highway:Lexus   -7.000000 -10.100144  -3.899856 0.0000738
highway:Subaru-city:Subaru      6.000000   2.899856   9.100144 0.0003268
city:Toyota-city:Subaru       -13.000000 -16.100144  -9.899856 0.0000001
highway:Toyota-city:Subaru     -8.000000 -11.100144  -4.899856 0.0000190
city:Toyota-highway:Subaru    -19.000000 -22.100144 -15.899856 0.0000000
highway:Toyota-highway:Subaru -14.000000 -17.100144 -10.899856 0.0000000
highway:Toyota-city:Toyota      5.000000   1.899856   8.100144 0.0016653
```

Two-way ANOVA syntax.

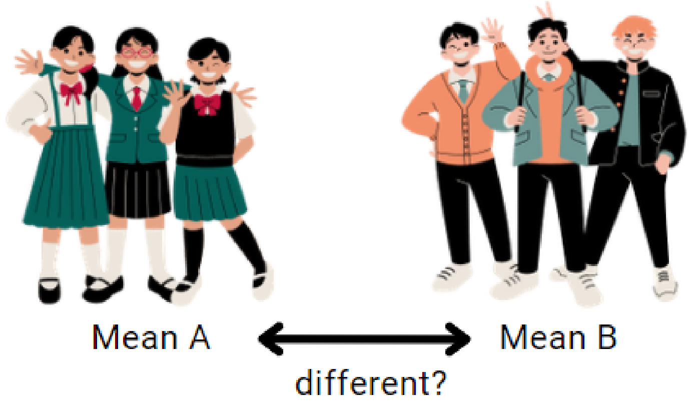
Furthermore, the **Tukey test** helps us figure out where *the differences* are lying the most, which specific groups' means are different. It compares all possible pairs of means (every single one of them).
The **ggplot graph** below also helps us understand the results better!

```
graphMPG <- Twoway%>%
  group_by(where2, brand)%>%
  summarize(aveMPG2 = mean(mpg2), se = sd(mpg2) / sqrt(length(mpg2)), tstar =
qt(1-0.05/2,length(mpg2)-1))%>%
  ggplot(aes(x = where2,y = aveMPG2)) + geom_point(aes(color = brand),size=3)
+geom_errorbar(aes(ymin=aveMPG2-se * tstar, ymax=aveMPG2 +se*tstar),width=0.1)+
ylab("Average MPG")

library(plotly)
ggplotly(graphMPG)
```

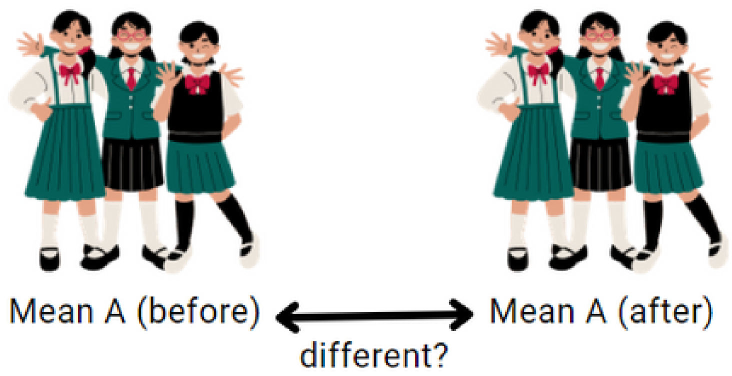


Key Takeaways:

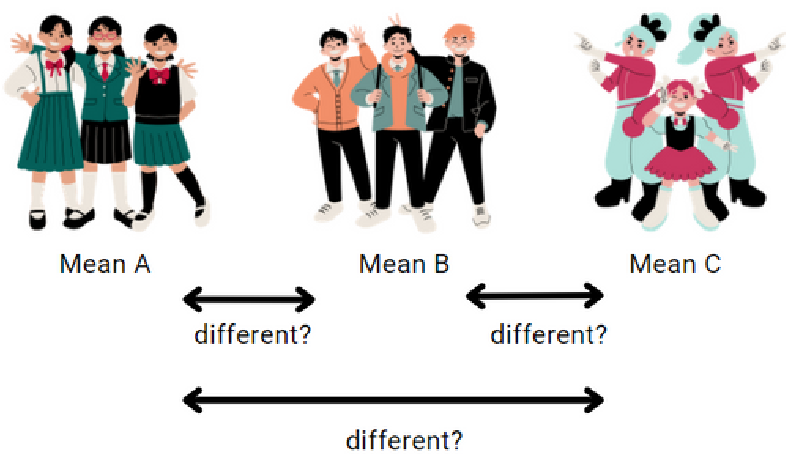
Independent t-test: if samples are from two populations.



Paired t-test: if samples are from one population, useful in the "before-after" scenario.



One-way ANOVA: compare means for more than two groups.



Two-way ANOVA: compare means for each factor and test if there is an interaction between factors for more than two groups.