/BIOSTATS 4
/ SUMMER RESEARCH

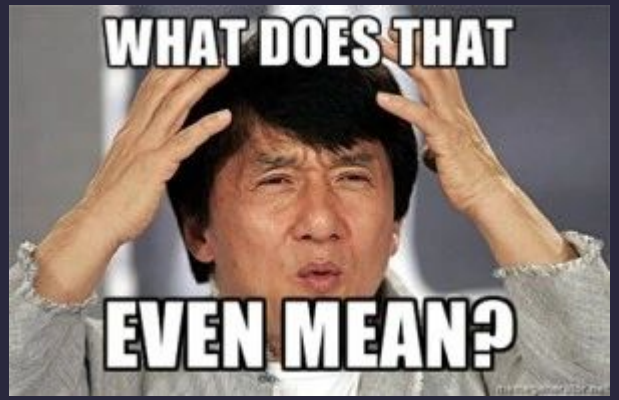
Coding with Real Applications for
Students in Healthcare


You gotta do the cooking by the book
You know you can't be lazy

/Choosing a Test

Three criteria are decisive for the selection of the statistical test, which are as follows:

- the number of variables,
- types of data/level of measurement (continuous, binary, categorical) and
- the type of study design (paired or unpaired).



/Number of Variables

- Single Variables = Descriptive Statistics
  a. Mean , median , standard deviation , boxplot etc
- >1 Variables = Relationships
  a. Most of the “tests” ie t–Test, Chi Squared, Rank Sum Test

Descriptive and Inferential Statistics

cuemath
THE MATH EXPERT

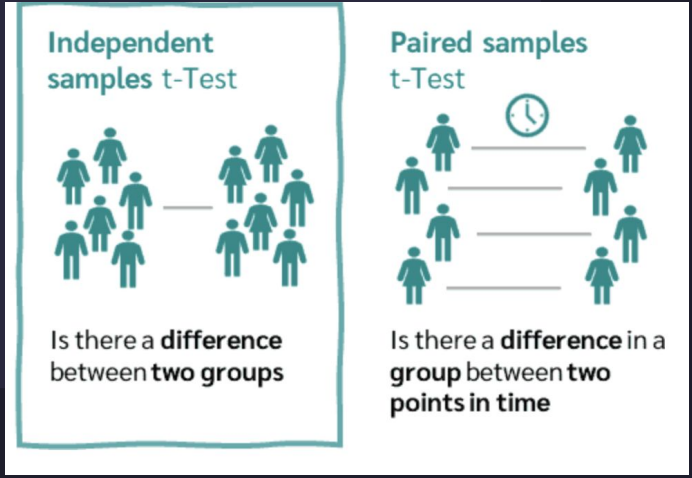
Descriptive Statistics		Inferential Statistics	
Measures of Central Tendency	Measures of Dispersion	Hypothesis Testing	Regression Analysis
Mean Median Mode	Range Standard Deviation Variance Absolute Deviation	Z test F test T test	Linear Regression

/Level of Measurement

- Continuous
  a. Can take any value within a certain range (infinite values between 0 and 1!)
  b. ex BP can be 120.1 , 101.7 etc mmHg
  c. Each individual number therefore has zero probability, we can only assign probability to ranges of values ie (P (BP < 115)
- Binary
  a. Yes or No
  b. Ex “Is BP > 125”
  c. Each level (Y/N) has a distinct probability
- Categorical
  a. Can take only enumerated values
  b. Ex BP is [“High”, “Low”, “Normal”]
  c. Each level has a distinct probability

/Type of Study Design

- Paired = Dependent
  a. results can be obtained for each patient under all experimental conditions
  b. Ex- Measure **ONE** person's heart rate before and after running a marathon
  c. Ex- Measure **ONE** person's eczema levels on Drug A and on Drug B (each on different arms)
- Unpaired = Independent
  a. results for each patient are only available under a single set of conditions
  b. Ex
     i. Group A all take drug A (only)
     ii. Group B all take drug B (only)
     iii. We then measure the difference between **GROUPS**

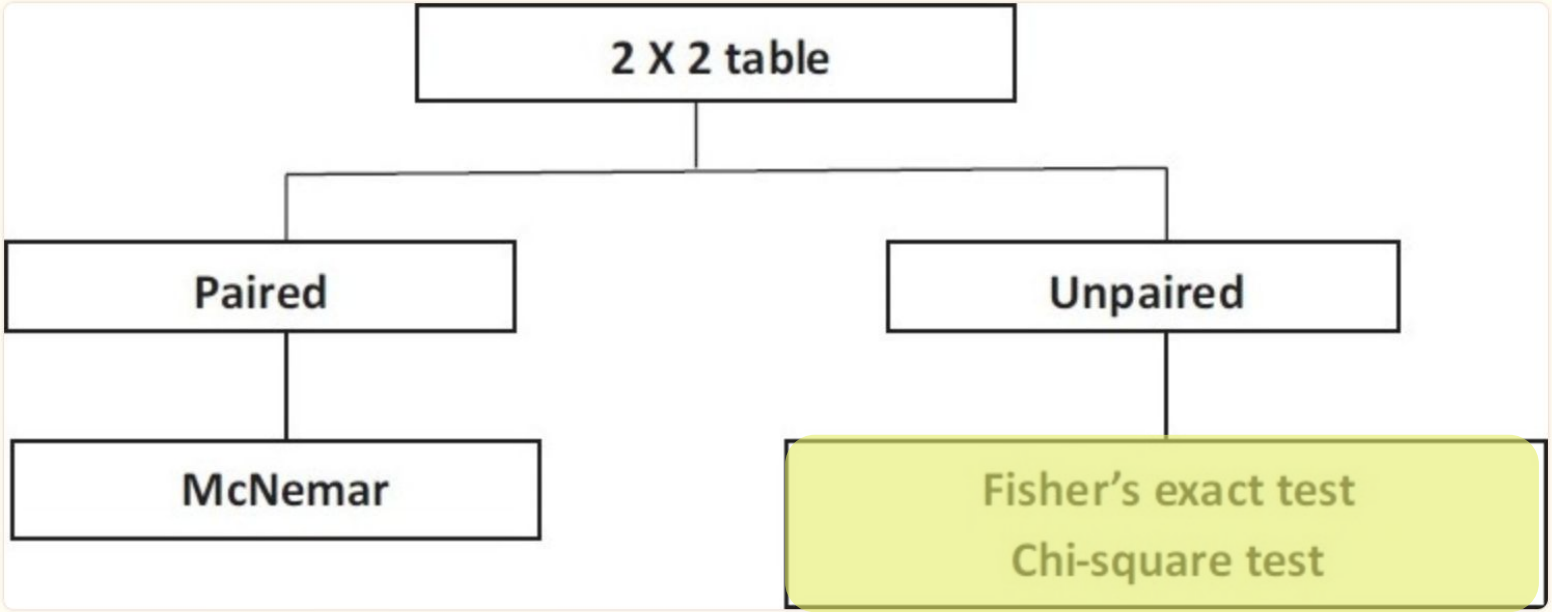


These 3 things are all you need to know to choose statistical tests!

You want to study Sweetener's Effect on Bladder Cancer

	Bladder Cancer	
Sweetener	Yes	No
Used	129	245
Never Used	171	332

- the number of variables =
- types of data/level of measurement
  a. Sweetener =
  b. Cancer =
- the type of study design
  a.


2 X 2 table
Paired
Unpaired
McNemar
Fisher's exact test
Chi-square test

Fishers's Exact Test

```r
dat <- data.frame(
  "smoke_no" = c(7, 0),
  "smoke_yes" = c(2, 5),
  row.names = c("Athlete", "Non-athlete"),
  stringsAsFactors = FALSE
)
colnames(dat) <- c("Non-smoker", "Smoker")

dat
```

```
##             Non-smoker Smoker
## Athlete              7      2
## Non-athlete          0      5
```

Fisher's exact test in R

To perform the Fisher's exact test in R, use the `fisher.test()` function as you would do for the Chi-square test:

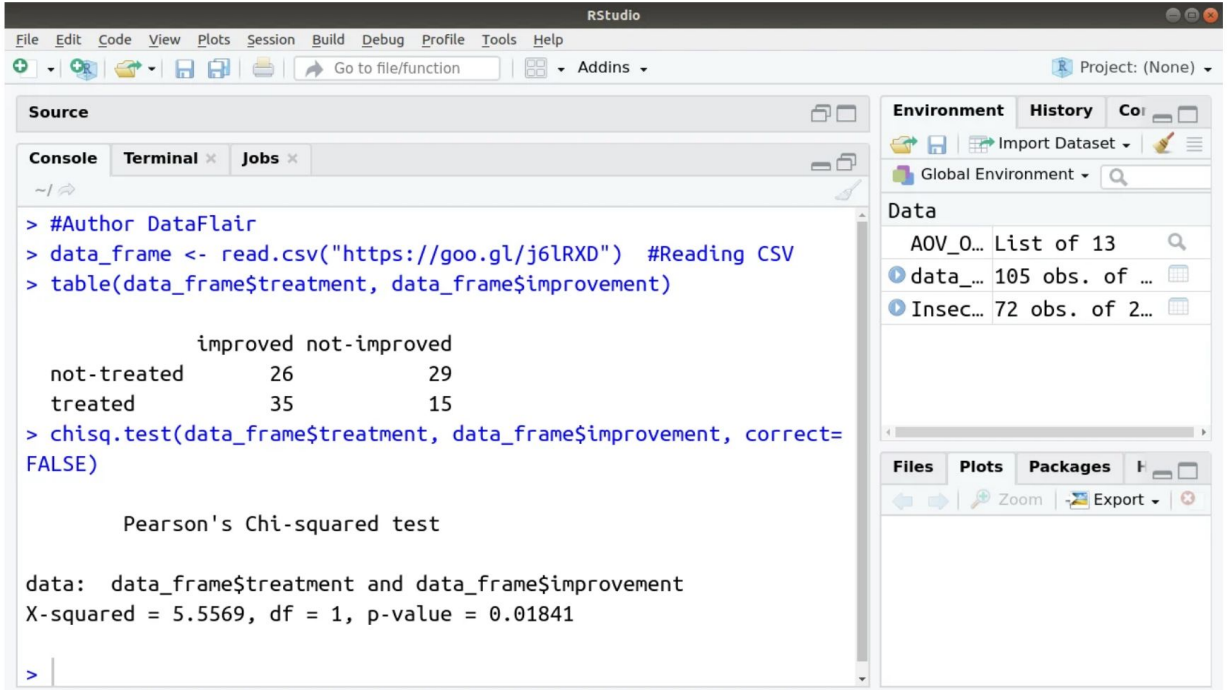
```r
test <- fisher.test(dat)
test
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  dat
## p-value = 0.02098
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.449481      Inf
## sample estimates:
## odds ratio
##        Inf
```

⇒ In our context, rejecting the null hypothesis for the Fisher's exact test of independence means that there is a significant relationship between the two categorical variables (smoking habits and being an athlete or not). Therefore, knowing the value of one variable helps to predict the value of the other variable.

Chi Square Test

Output:

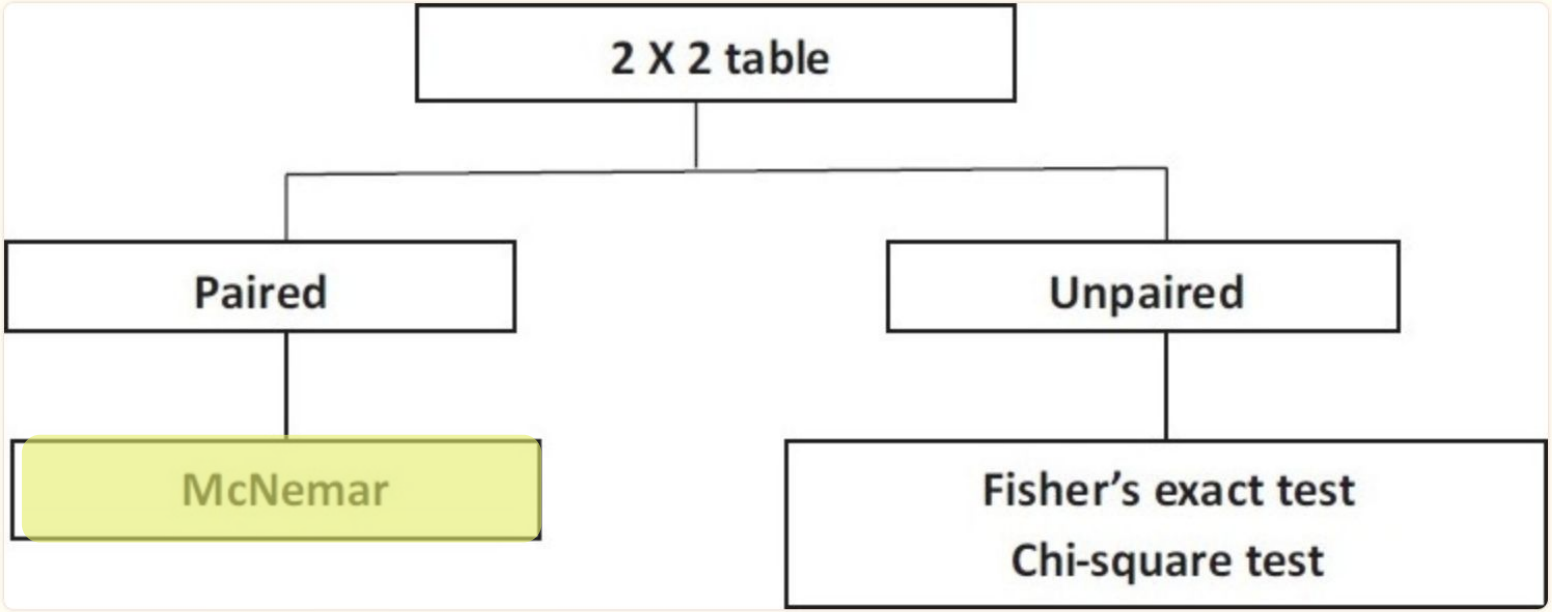


We have a chi-squared value of 5.5569. Since we get a p-Value less than the significance level of 0.05, we reject the null hypothesis and conclude that the two variables are in fact dependent.

You want to study joint pain before and after treatment

			Experienced Joint Pain After Treatment		Total
			No	**Yes**	
Experienced Joint Pain Before Treatment	**No**	Count	215	75	290
	Yes	Count	785	380	1165
Total		Count	1000	455	1455

- the number of variables =
- types of data/level of measurement
  a. Before =
  b. After =
- the type of study design
  a.


2 X 2 table
Paired
Unpaired
McNemar
Fisher's exact test
Chi-square test

McNemar Test

```
data

                 Before Video
After Video      Support Do Not Support
  Support             30             40
  Do Not Support      12             18
```

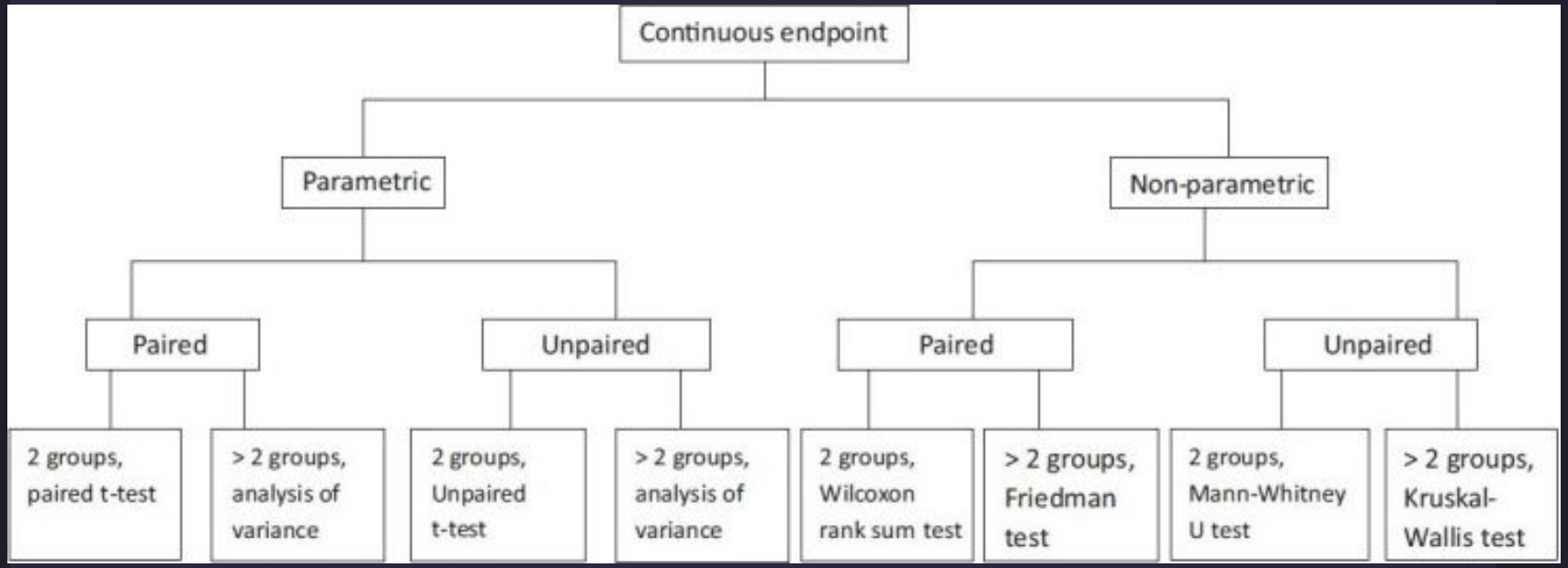
```
#Perform McNemar's Test with continuity correction
mcnemar.test(data)

        McNemar's Chi-squared test with continuity correction

data:  data
McNemar's chi-squared = 14.019, df = 1, p-value = 0.000181

#Perform McNemar's Test without continuity correction
mcnemar.test(data, correct=FALSE)

        McNemar's Chi-squared test

data:  data
McNemar's chi-squared = 15.077, df = 1, p-value = 0.0001032
```

In both cases the p-value of the test is less than 0.05, so we would reject the null hypothesis and conclude that the proportion of people who supported the law before and after watching the marketing video was statistically significant different.

A Note on Packages

- Packages are collections of modules (AKA libraries) that organize code and make it easier to use
  - Typically contain pre-written **functions**, classes, and variables
- A very important one for R is **dplyr**, which is a grammar framework for data manipulation
  - Contains the statistical tests with the syntax previously shown, plus others
  - If dplyr is is not already installed, it can be with the *install.packages( )* command and loaded in normal fashion with *library( )*

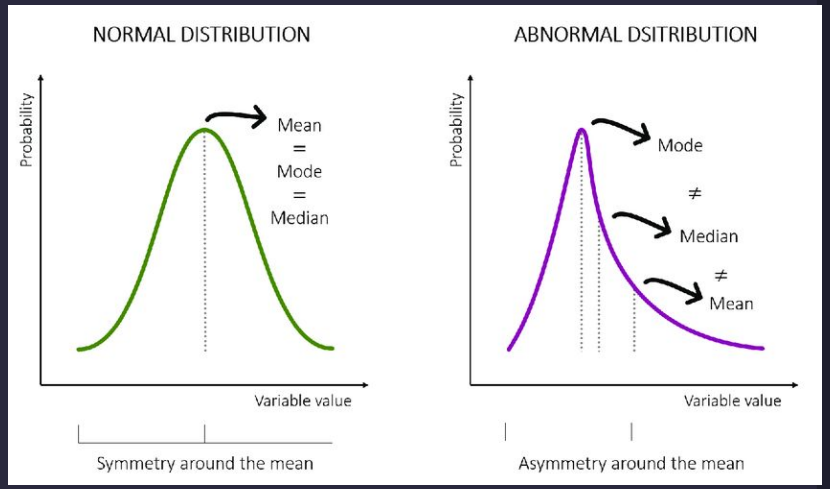
/Continuous Data


Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test


Mom come
pick me up
I'm scared

/Okay we lied on more thing

- Parametric
  a. Endpoint is normally distributed
- Non- Parametric
  a. Endpoint is not normally distributed
- How can I tell?
  a. Shapiro Wilk Test or Kolmogorov Test



Shapiro Wilk Test

Syntax:

shapiro.test(x)

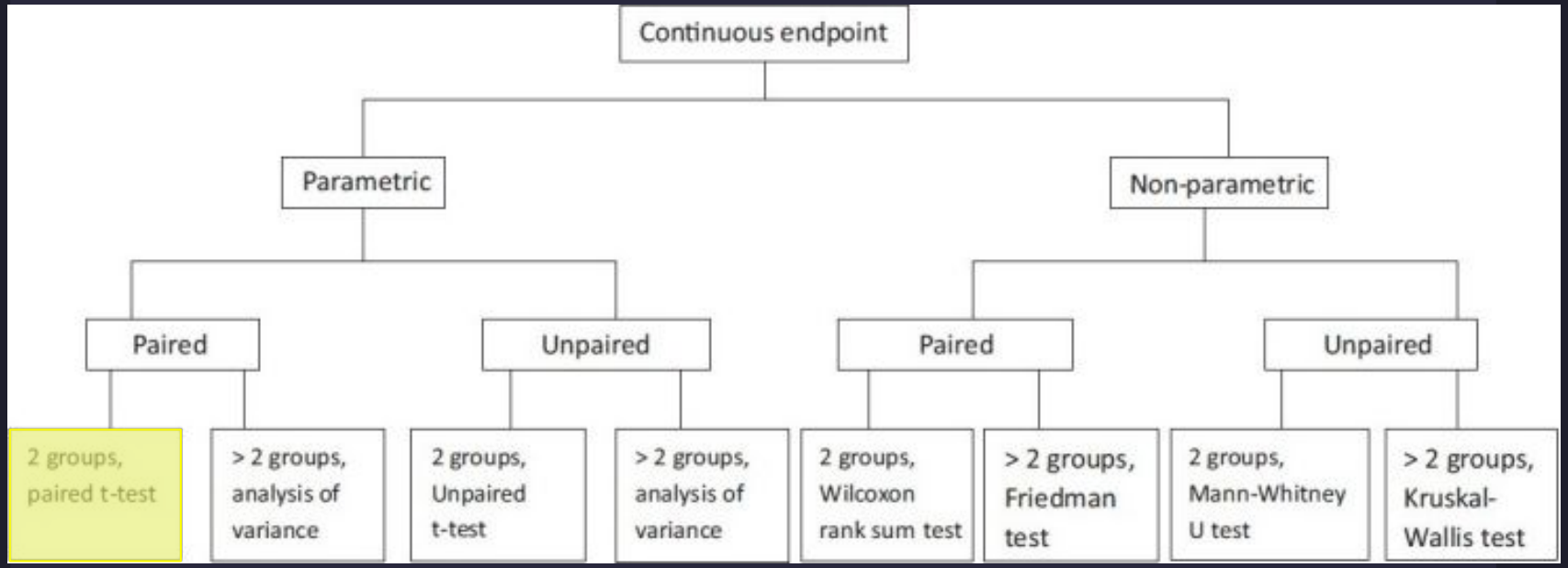
Parameter:

x *: a numeric vector containing the data values. It allows missing values but the number of missing values should be of the range 3 to 5000.*

```
> shapiro.test(my_data$len)

    Shapiro-Wilk normality test

data:  my_data$len
W = 0.96743, p-value = 0.1091
```

- Null hypothesis is that data ***is*** normally distributed
- $p < 0.05$ suggests significant difference, reject null (i.e., data is not normally distributed)
- $p > 0.05$ suggests no significant difference, accept null

From the output obtained we can assume normality. The p-value is greater than 0.05. Hence, the distribution of the given data is not different from normal distribution significantly.

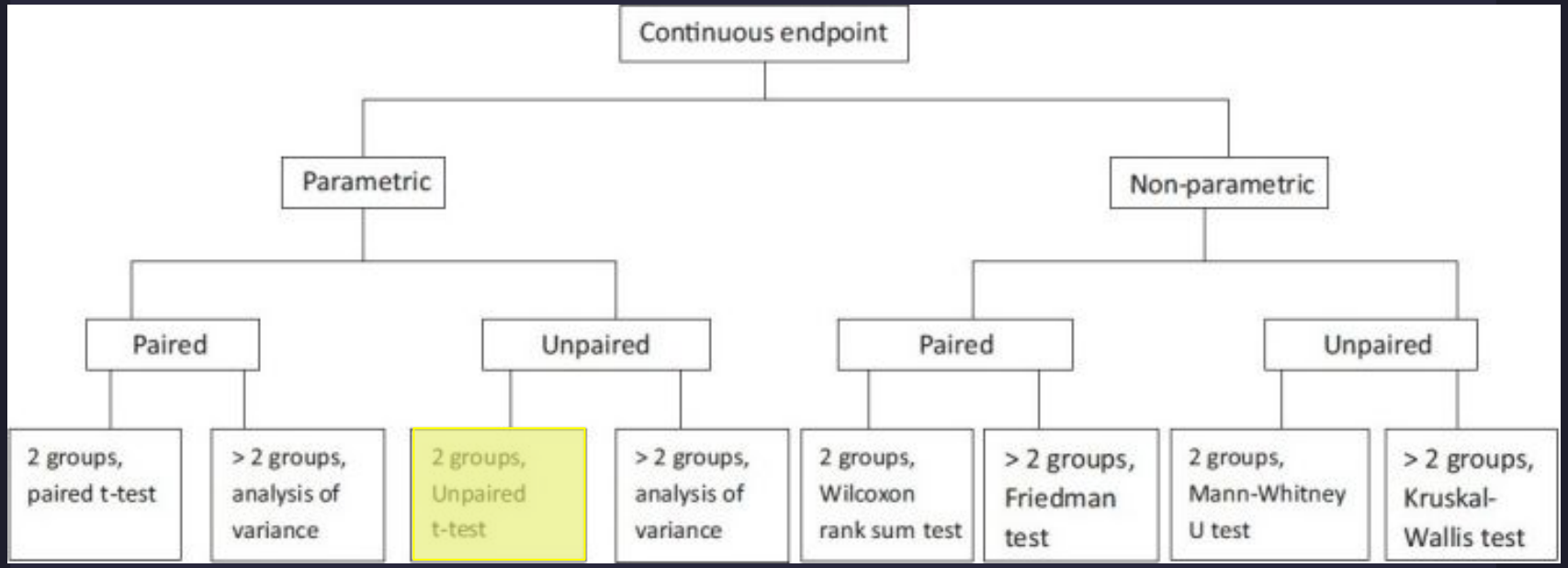

Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

Paired T Test

```r
sweetOne <- c(rnorm(100, mean = 14, sd = 0.3))
sweetTwo <- c(rnorm(100, mean = 13, sd = 0.2))

t.test(sweetOne, sweetTwo, paired = TRUE)
```

```
    Paired t-test

data:  sweetOne and sweetTwo
t = 29.31, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.9892738 1.1329434
sample estimates:
mean difference
       1.061109
```

- Data comes from measurements of an <u>identical group/set of observations</u> made under two different conditions (ex. different time)
- Null hypothesis is that the difference between the means is *not noticeably different from zero*
- $p < 0.05$ means there is a significant difference
- $p > 0.05$ means there is not a significant difference

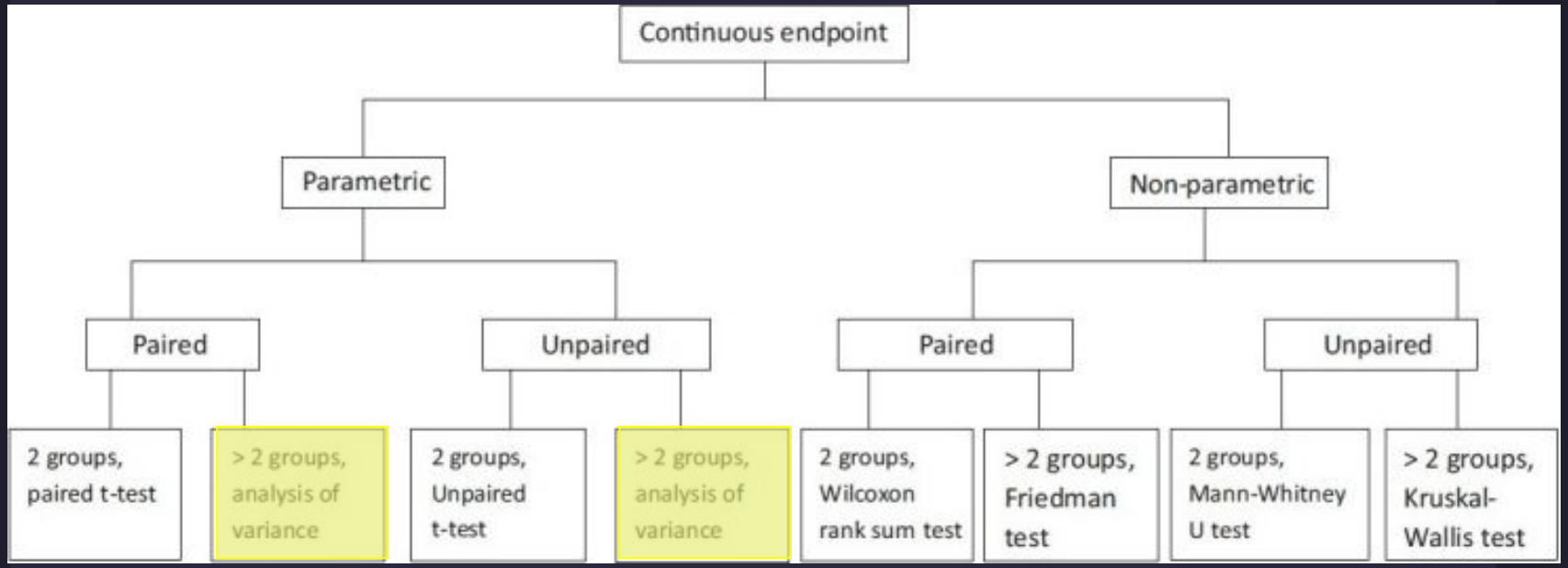

Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

Unpaired T Test

```r
shopOne <- rnorm(50, mean = 140, sd = 4.5)
shopTwo <- rnorm(50, mean = 150, sd = 4)

t.test(shopOne, shopTwo, var.equal = TRUE)
```

```
    Two Sample t-test

data:  shopOne and shopTwo
t = -13.158, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.482807  -8.473061
sample estimates:
mean of x mean of y
 140.1077  150.0856
```

- Null hypothesis is that there is **no significant difference between the means** of the sample populations
- Note: *var.equal = TRUE* is a flag to prevent R from running Welch's test, since R assumes variances are unequal if not specified.


Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

ANOVA (One-way)

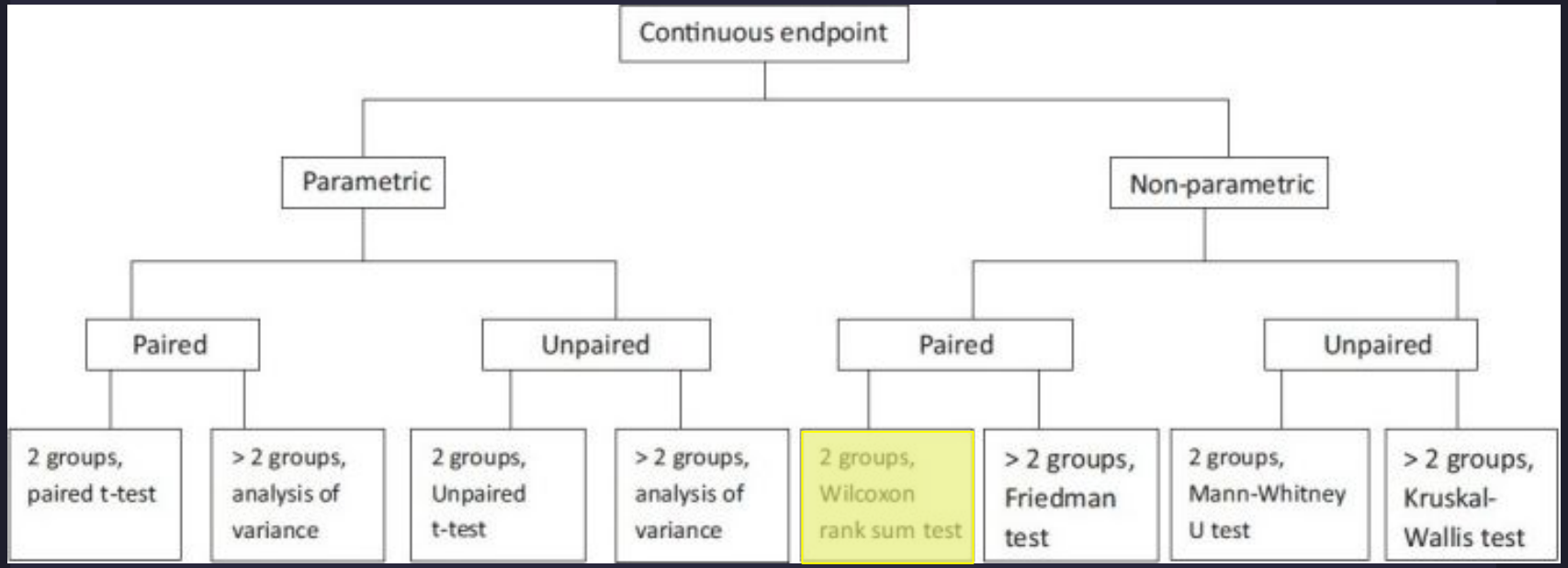
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

← data set

```r
mtcars_aov <- aov(mtcars$disp~factor(mtcars$gear))
summary(mtcars_aov)
```

```
                    Df Sum Sq Mean Sq F value   Pr(>F)
factor(mtcars$gear)  2 280221  140110   20.73 2.56e-06 ***
Residuals           29 195964    6757
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Null hypothesis is that means of *all* categories are the same
- *aov* function does test and assigns results to a variable
- In this example, mean of “gears” category is significantly different

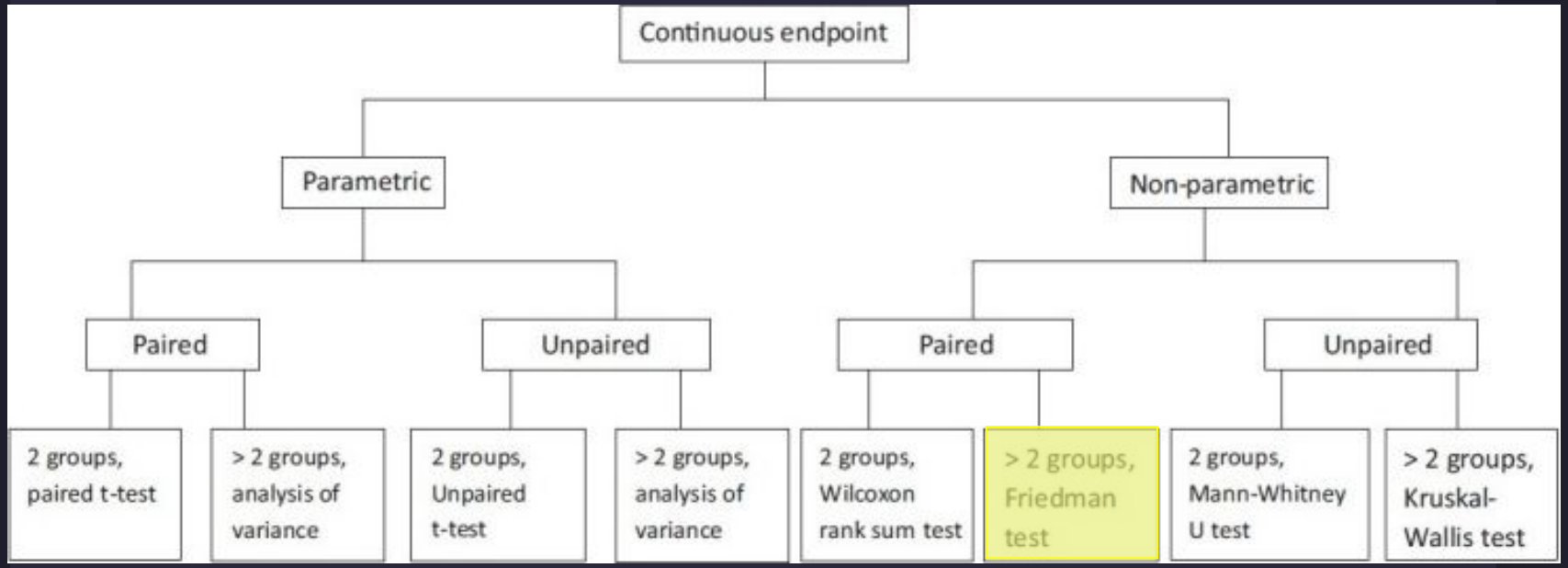

Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

Rank Sum Test

```r
# Paired Samples Wilcoxon Test

# The data set
# Weight of the rabbit before treatment
before <-c(190.1, 190.9, 172.7, 213, 231.4,
        196.9, 172.2, 285.5, 225.2, 113.7)

# Weight of the rabbit after treatment
after <-c(392.9, 313.2, 345.1, 393, 434,
        227.9, 422, 383.9, 392.3, 352.2)

# Create a data frame
myData <- data.frame(
group = rep(c("before", "after"), each = 10),
weight = c(before, after)
)

# Print all data
print(myData)

# Paired Samples Wilcoxon Test
result = wilcox.test(before, after, paired = TRUE)

# Printing the results
print(result)
```

```
    Wilcoxon signed rank test

data:  before and after
V = 0, p-value = 0.001953
alternative hypothesis: true location shift is not equal to 0
```

- Null hypothesis is that median weights of population are not significantly different between observations
- Since $p < 0.05$ in the example, we reject the null


Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

A Note on Linear Algebra (Scary!)

- When coding, you will frequently see the terms vector and matrix used in (potentially) unfamiliar contexts. Understanding what is meant by is often important for proper syntax.
- Scalars are single numbers
  - Ex. 5.5 is a scalar quantity
- Vectors are lists of numbers arranged along an axis
  - Ex. **v** = {1, 2, 3, 4} is a vector
  - All of the previous example data sets were entered into R as vectors
- Matrices are numbers arranged along two axes
  - Ex.

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \\ M_{31} & M_{32} \end{bmatrix}$$

- The above are 0th, 1st, and 2nd order tensors, respectively. Tensors are generalizations of the above. You are unlikely to encounter higher order tensors, but may find the term used in specific applications.

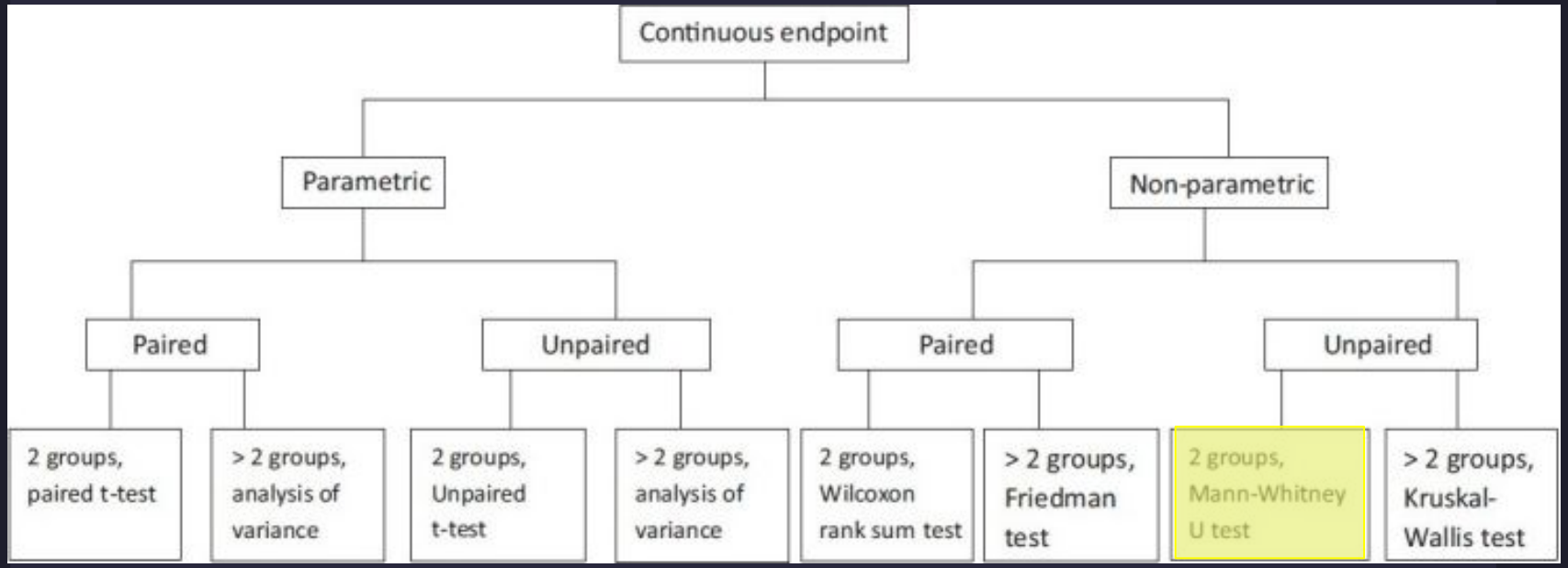
Friedman Test

```r
# Friedman Test

#input the data
y <- matrix(c(1.24,1.50,1.62,
              1.71,1.85,2.05,
              1.37,2.12,1.68,
              2.53,1.87,2.62,
              1.23,1.34,1.51,
              1.94,2.33,2.86,
              1.72,1.43,2.86),
nrow = 7, byrow = TRUE,
dimnames = list(Person= as.character(1:7),Drugs = c("Drug A","Drug B","Drug C"))
```

```r
result = friedman.test(y)
print(result)
```

```
    Friedman rank sum test

data:  y
Friedman chi-squared = 8.8571, df = 2, p-value = 0.01193
```

- Null hypothesis is that all three groups (drugs) have the same probability distribution


Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

Mann-Whitney test

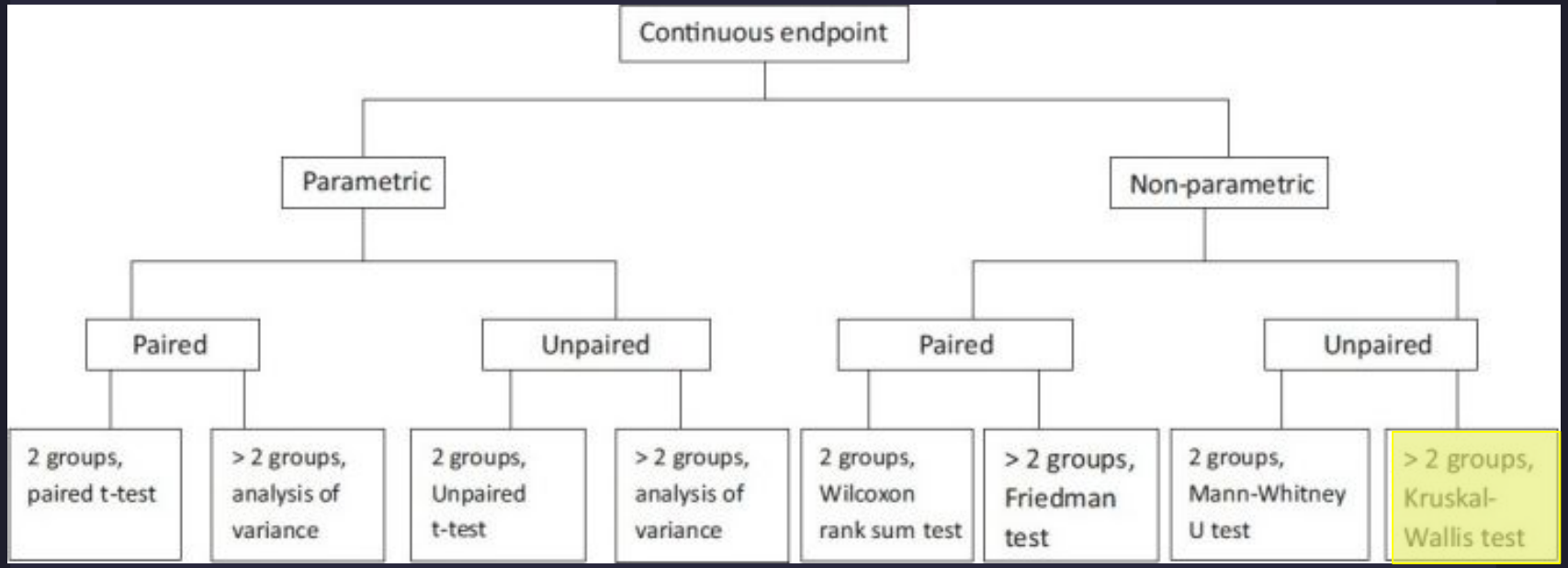
```r
# Creating a vector of red bulb and orange prices
red_bulb <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8)
orange_bulb <- c(47.8, 60, 63.4, 76, 89.4, 67.3, 61.3, 62.4)

# Passing them in the columns
BULB_PRICE = c(red_bulb, orange_bulb)
BULB_TYPE = rep(c("red", "orange"), each = 8)

# Now creating a dataframe
DATASET <- data.frame(BULB_TYPE, BULB_PRICE, stringsAsFactors = TRUE)
```

```r
res <- wilcox.test(BULB_PRICE~ BULB_TYPE,
                   data = DATASET,
                   exact = FALSE)
res
```

	BULB_TYPE	BULB_PRICE
1	red	38.9
2	red	61.2
3	red	73.3
4	red	21.8
5	red	63.4
6	red	64.6
7	red	48.4
8	red	48.8
9	orange	47.8
10	orange	60.0
11	orange	63.4
12	orange	76.0
13	orange	89.4
14	orange	67.3
15	orange	61.3
16	orange	62.4

```
        Wilcoxon rank sum test with continuity correction

data:  BULB_PRICE by BULB_TYPE
W = 44.5, p-value = 0.2072
alternative hypothesis: true location shift is not equal to 0
```

- Null hypothesis is that the two ordinal categories have the same distribution
- Box plot is particularly helpful here for visualization of distributions, requires separate packages to generate


Continuous endpoint
Parametric
Non-parametric
Paired
Unpaired
Paired
Unpaired
2 groups, paired t-test
> 2 groups, analysis of variance
2 groups, Unpaired t-test
> 2 groups, analysis of variance
2 groups, Wilcoxon rank sum test
> 2 groups, Friedman test
2 groups, Mann-Whitney U test
> 2 groups, Kruskal-Wallis test

Kruskal-Wallis Test

```r
# Kruskal-Wallis Test

# Taking the PlantGrowth data set
myData = PlantGrowth

# Performing Kruskal-Wallis test
result = kruskal.test(weight ~ group,
                      data = myData)
print(result)
```

```
    weight group
1     4.17  ctrl
2     5.58  ctrl
3     5.18  ctrl
4     6.11  ctrl
5     4.50  ctrl
6     4.61  ctrl
7     5.17  ctrl
8     4.53  ctrl
9     5.33  ctrl
10    5.14  ctrl
11    4.81  trt1
12    4.17  trt1
13    4.41  trt1
14    3.59  trt1
15    5.87  trt1
16    3.83  trt1
17    6.03  trt1
18    4.89  trt1
19    4.32  trt1
20    4.69  trt1
21    6.31  trt2
22    5.12  trt2
23    5.54  trt2
24    5.50  trt2
25    5.37  trt2
26    5.29  trt2
27    4.92  trt2
28    6.15  trt2
29    5.80  trt2
30    5.26  trt2
[1] "ctrl" "trt1" "trt2"
```

Kruskal-Wallis rank sum test

data: weight by group

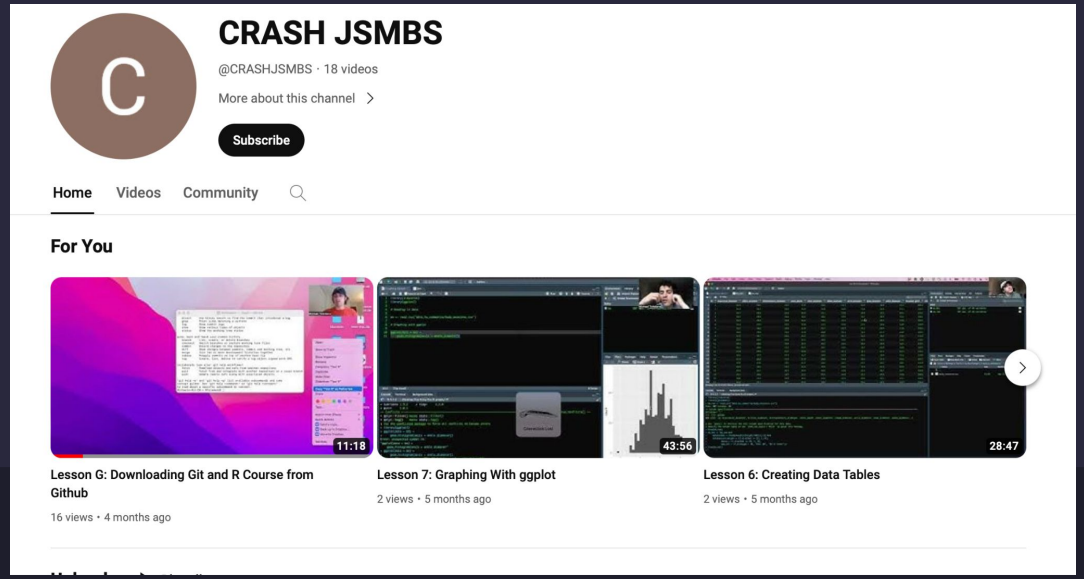
Kruskal-Wallis chi-squared = 7.9882, df = 2, p-value = 0.01842

- Null hypothesis is that there are no significant differences between the distributions of all groups
- Note: does not specify which group differs - best determined visually via box plot

/How do I run tests?

- Luckily a member of previous Eboard, Mikey Toledano, and our current pres Anneliese Markus made an entire course on how to conduct data analysis in R, including youtube videos!
- And we will be hosting further sessions on coding for data analysis and figure generation

https://github.com/jsmbsCRASH/CRASH_R



```
if(Question == TRUE) {print("I don't know!")}
```