Inferential Statistics is where the sample draws conclusion about the population!

One Sample T-Test is a statistical hypothesis used to determine whether an unknown population mean is different from a specific value.
Two Sample T-Test is statistical hypothesis used to determine if two population means are equal.
When means of Two-Samples are compared with eachother, then it is considered Two Sample Independent T-test.
When the same subjects are paired or matched in some way, then it is considered as Two Sampled Paired Dependent T-test.
Remember that for a One Sample and a Two Sample T-test, we often use Two-Tailed T-test
One Sample = N-1 and Two Sample = N1+N2-2

One Sample Null Hypothesis is a statistical hypothesis that states that there is no statistical difference between the population mean and a specific value.
One Sample Alternative Hypothesis is a statistical hypothesis that states that there is a statistical difference between the population mean and a specific value.
Two Sample Null Hypothesis is a statistical hypothesis that states that there is no statistical difference between the means of the two populations.
Two Sample Alternative Hypothesis is a statistical hypothesis that states that there is a statistical difference between the means of the two populations.

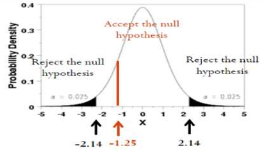
```
towns <-
c(7540,8421,8560,7412,8953,7859,6098)
res.ttest <-
t.test(towns,conf.level=0.99)
res.ttest
```

```
j    One Sample t-test
data:  towns
t = 21.845, df = 6, p-value = 6.012e-
07
alternative hypothesis: true mean is
not equal to 0
99 percent confidence interval:
 6505.022 9164.407
sample estimates:
mean of x
 7834.714
```

```
athlete <- data.frame(before=c(3.7,3.1,……… lang dataliste
forkortet),
after=c(4.0,3.1,………… lang dataliste forkortet))
res.ttest <-
t.test(athlete$before,athlete$after,paired=TRUE)
res.ttest
```

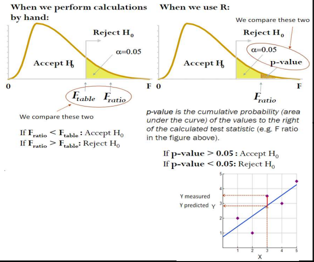
```
> res.ttest
    Paired t-test
data:  athlete$before and athlete$after
t = -2.997, df = 19, p-value = 0.007411
alternative hypothesis: true mean difference is not equal
to 0
95 percent confidence interval:
 -0.24626407 -0.04373593
sample estimates:
mean difference
         -0.145
```

Library(agricolae)
Library(tidyr)
Library(readxl)
Library(readr)
Library(effectsize)
Library(modeest)
Library(rearrr)



If you want to see, if the data is normally distributed then use the shapiro.test(data$column), if p is greater than 0,05=95% then the data is normally distributed.
Type 1 Error: Rejecting a true Null Hypothesis.
Type 2 Error: Accepting a false Null Hypothesis.

One WAY ANOVA also known as Multisample T-test where the mean of three groups is compared.
Two WAY ANOVA is where we study the effect of 2 factors (treatments) on the response variable. Two factor = 2 levels!



One Way Null Hypothesis is a statistical hypothesis that states that there is no statistical difference between the means of the two populations.
One Way Alternative Hypothesis is a statistical hypothesis that states that there is a statistical difference between the means of the two populations.

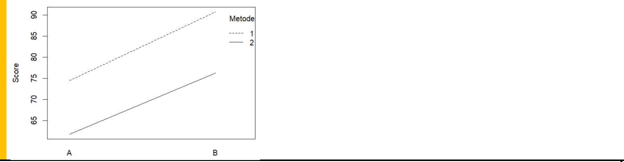
Critical Values and T-table
One Way ANOVA has d.f. of (k-1, N-k) and k means group and N means the amount of data in total of the groups combined.
One Sample T-test has d.f. of N-1 and N means the total amount of the individual sample size.
Two Sample T-test has d.f. of (N1+N2)-2 where N1 means the sample size of one sample group and N2 means the sample size of the other sample group.

Residuals
There is a difference between the location of a specific point in the scatter plot (Y measured), and what the regression equation predicts (Y predicted).
The size of these residuals provides information about the existence of other variables (variables other than X), that also affect the Y (and that have not been included in the regression analysis)

```
data <-
c(12,8,11,17,16,15,5,9,2,0,1,3,6,10,5,9,8,6)
treatment <-
c(rep("m",6),rep("e",6),rep("d",6))
methods <- data.frame(data,treatment)
str(methods)
methods$treatment <-
as.factor(methods$treatment)
res.aov <- aov(data~treatment,data=methods)
summary(res.aov)
print(LSD.test(res.aov,"treatment"))
```

```
> summary(res.aov)
            Df  Sum Sq  Mean Sq  F value
Pr(>F)
treatment    2  293.4   146.72   16.74
$groups
       data groups
m 13.166667      a
d  7.333333      b
e  3.333333      c
```

```
Uni <- data.frame(Metode=c("1",…),Lektor=c("A",….),Score=c(67,68,69…))
Uni$Metode <- as.factor(Uni$Metode)
Uni$Lektor <- as.factor(Uni$Lektor)
Two_way_anova_Uni <- aov(Score~Metode+Lektor+Metode*Lektor,data=Uni)
summary(Two_way_anova_Uni)
interaction.plot(Uni$Lektor,Uni$Metode,Uni$Score,xlab="Lektor",ylab="Score",tr
```



Descriptive Statistics
Used to describe, organize, summarize info.

Measures of Central Tendency (Location)
Measures of Location gives a description of a whole set of data with a single value that represents the middle or centre of its distribution.
Mean(), Median(), Mode and Midrange (m+s)/2

Null Hypothesis is where there is no relation between X and Y in the population. (Simple Regression)
Alternative Hypothesis is where there is a relation between X and Y in the population. (Simple Regression)

Measures of Variation (Dispersion)
Measures of Dispersion are concerned with the distribution of values around the mean in data.
Range(), Variance var() and Standard Deviation sd()

If Causation is mentioned in an article context, then it means that the researchers found that changes in one variable they measured, directly caused changes in the other.

Simple Regression Analysis
Is a method that describes how one independent predictor/explanatory variable x describes the output and response variable y.

$$y = a + \beta_1 \cdot x_1$$

Multiple Regression Analysis
Is a method that describes how more independent predictor/explanatory variable x describes the output and response variable y.

$$y = a + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_k \cdot x_k$$

Regression Analysis
Is a statistical method used to describe how one (or more) quantitative variable is related to another quantitative variable.

```
Drikke <- data.frame(Number=c(1,2,3,4,5),Blood=c(10,8,12,16,20))
cor.test(Drikke$Number,Drikke$Blood)
plot(Drikke$Number,Drikke$Blood)
regression_Drikke <- lm(Blood~Number,data=Drikke)
abline(regression_Drikke)
regression_Drikke
predict(regression_Drikke,data.frame(Number=4.2),interval="predict")
```

```
> cor.test(Drikke$Number,Drikke$Blood)
data:  Drikke$Number and Drikke$Blood
t = 4.0415, df = 3, p-value = 0.02726
sample estimates:
     cor
0.919145
> regression_Drikke
Call:
lm(formula = Blood ~ Number, data = Drikke)
Coefficients:
(Intercept)       Number
        4.8          2.8
> predict(regression_Drikke,data.frame(Number=4.2),interval="predict")
   fit      lwr      upr
1 16.56 8.476837 24.64316
```

```
regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
summary(regression_Healthcare)

(Multiple Linear Regression)
```

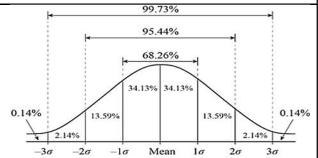
```
> regression_Healthcare <- lm(TREATCOST~CAREDAYS,data=Healthcare)
> summary(regression_Healthcare)
Call:
lm(formula = TREATCOST ~ CAREDAYS, data = Healthcare)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     6463       5927   1.091    0.277
CAREDAYS       16572       1169  14.174   <2e-16 ***

Residual standard error: 35460 on 136 degrees of freedom
Multiple R-squared:  0.5963,    Adjusted R-squared:  0.5933
F-statistic: 200.9 on 1 and 136 DF,  p-value: < 2.2e-16
```

Measures of Position
Measures of position indicate the position of a value, relative to other values in a set of observations.
Percentile, Decile and Quartile and Outlier
*Percentile is selfish (before/amount) *100*

$$L_k = \frac{percentile}{100} \cdot (n+1)$$
$$Q1 - 1{,}5 \cdot (IQR)$$
$$Q3 + 1{,}5 \cdot (IQR)$$

Normal Distribution
A symmetric distribution defined on the range -∞ to +∞ whose shape is defined by two parameters, the mean, denoted μ, that centers the distribution, and the standard deviation, σ, that determines the spread of the distribution.
68% of area under the curve between $\mu \pm \sigma$
95% of area under the curve between $\mu \pm 2\sigma$
99,7% of area under the curve between $\mu \pm 3\sigma$



Distributions & Probabilities
A Probability Experiment is a chance process that leads to well-defined results called outcomes

Binomial Distribution
The experiment consists of n identical trials (simple experiments).
Each trial results in one of two outcomes (success or failures).
The probability of success on a single independent trial is equal to pi and pi remains the same from trial to trial.

$$P(y) = \frac{n!}{y! \cdot (n-y)} \cdot \pi^{y}(1-\pi)^{n-y}$$

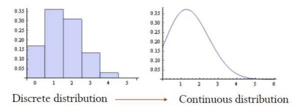
A survey found that one out of five Americans.
say he or she has visited a doctor in any given month.
If 10 people are selected at random, find the probability.
that exactly 3 will have visited a doctor last month.

$$\frac{10!}{3! \cdot (10-3)!} \cdot \left(\frac{1}{5}\right)^3 \cdot \left(\frac{4}{5}\right)^7 = \frac{393216}{1953125} \approx 0{,}2013266 \cdot 100\% = 20{,}132\%$$

Der er 20,132% sandsynlighed for at der var nøjagtig 3 der besøgte lægen sidst måned.

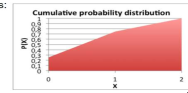
Types of Probability Distributions
A discrete random variable can assume a countable number of values.
- E.g., number of students present.

A continuous random variable can assume any value along a given interval.
- E.g., height of the students, the amount of sugar in an orange.



The cumulative probability distribution of X is:
P(X=0) + P(X=1) + P(X=2)

X	P(X=x)	P(X≤x)
0	1/4 = 0.25	0.25
1	2/4= 0.50	0.25+0.5=0.75
2	1/4 = 0.25	0.75+0.25=1



Cumulativ Probability Distribution $P(X \leq 1) = P(X=0) + P(X=1)$
Refers to the probability that the value of a random variable falls within a specified range.

Study Design
Process of planning experiments/data collection…….

Observational Study:
1. **Does NOT control the input variables.**
2. **Observes and identifies the natural relationships between the input variables and output variable.**

Experimental Study:
1. **The Investigator: Controls the input variables.**
2. **The Investigator: Observes and identifies the reasons for changes in the output variable.**

Skewness
Positive Right Skewness: When the mean is greater than the median.
Negative Left Skewness: When the median is greater than the mean.
No-Skewness: When the mean and median are coincided.
Kurtosis measures how peaked the histogram is
Leptokurtic (peaked), Mesokurtic (between), Platykurtic (flat)

Nice to Know
Least square difference interval used (after the analysis of variance has been completed), to make pairwise comparisons between means.
Correlation: Is statistical method used to describe the nature of the relationship between variables (eg. Positive or negative, linear, or non-linear). (Symbol: r)
Determination Coefficient: Is the ratio of the explained variation to the total variation. (Symbol: R^2)
If you want to make histogram, then use the hist(x=name_of_the_dataframe$numeric_column).
Combinatorics formular is only the center bracket which in the binomial distribution.

$$\frac{n!}{y! \cdot (n-y)}$$