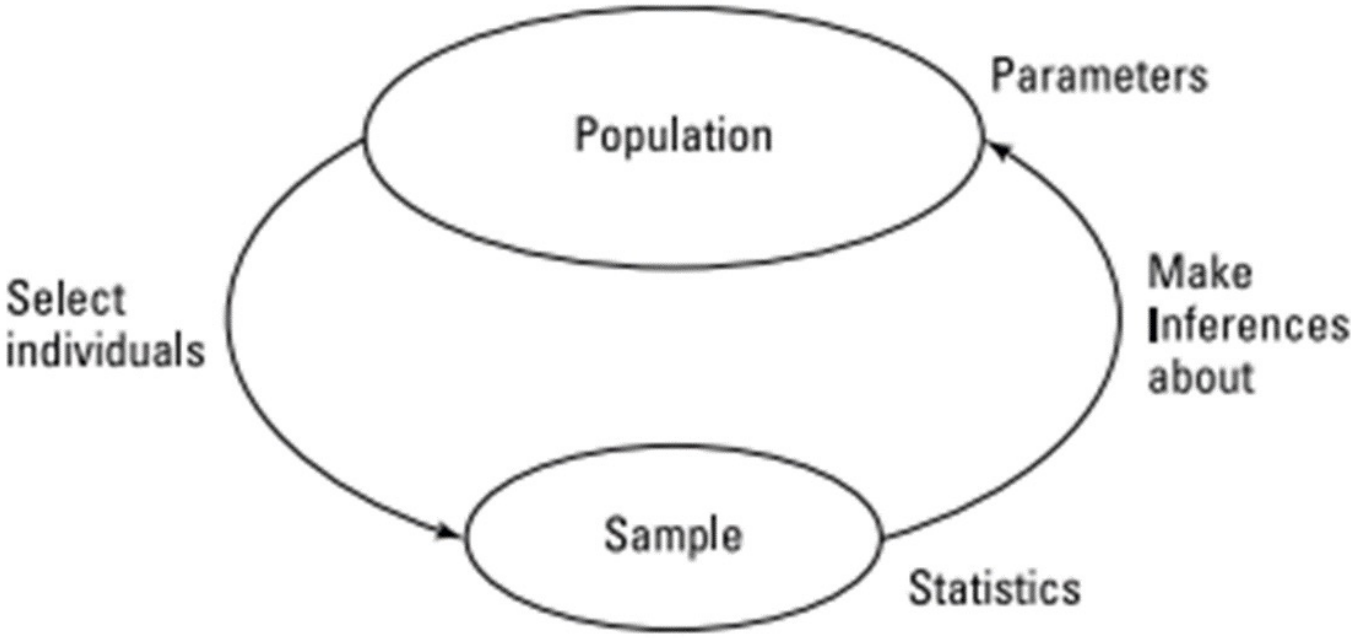
BASIC STATISTICS FOR DATA SCIENCE

INTRODUCTION

- Statistics help to make decision based on data
- What are the numbers saying ?
- Are there any trends in the data ?
- Can we make predictions from the data?
- Using data from samples to draw conclusions about the population
- The sample should be representative of the population
- Predictive Analytics
  - Regression – Linear regression,
    - Correlation coefficient
  - Multiple Linear Regression
  - Logistic Regression

CATEGORIES OF STATISTICS

- Descriptive
  - Organize data and focus on main characteristics of the data
  - Summary of data (Average, Mode, SD)
  - To describe the basic features of the data
- Inferential
  - Generalizes the larger dataset and applies probability theory to draw conclusion
  - Infer population parameters and model relationships within data
  - Modeling : dev mathematical equations which describes relationships between variables
  - Make generalization about the broader population



STATISTICAL TERMS

- Population
  - The group from which data is to be collected
- Sample
  - A subset of a population

STATISTICAL TERMS

- Variable
  - Feature characteristic of any member of a population differing in quality or quantity from another member. Eg. Age, race, height, weight, gender, degree….
- Quantitative variable
  - A variable differing in quantity eg, weight of a person, number of pax
- Qualitative Variable
  - A variable differing in quality eg. Color, degree of damage to a car in an accident
- Discrete Variable
  - One which no value can be assumed between two given values
- Continuous variable

STATISTICAL TERMS

- Independent variable
- Dependent variable –what is measured
- Assess the relationship between them
  - to find out whether changes in an independent variable are associated with changes in a dependent variable
- Types of data in statistics
  - Nominal
  - Ordinal
  - Interval
  - Ratio

STATISTICAL MEASURES

Measures of Frequency | Measures of Central Tendency | Measures of Spread | Measures of Position

Frequency of the data indicates the number of occurrences of any particular data value in the given dataset.

The measures of frequency are number and percentage.

STATISTICAL MEASURES

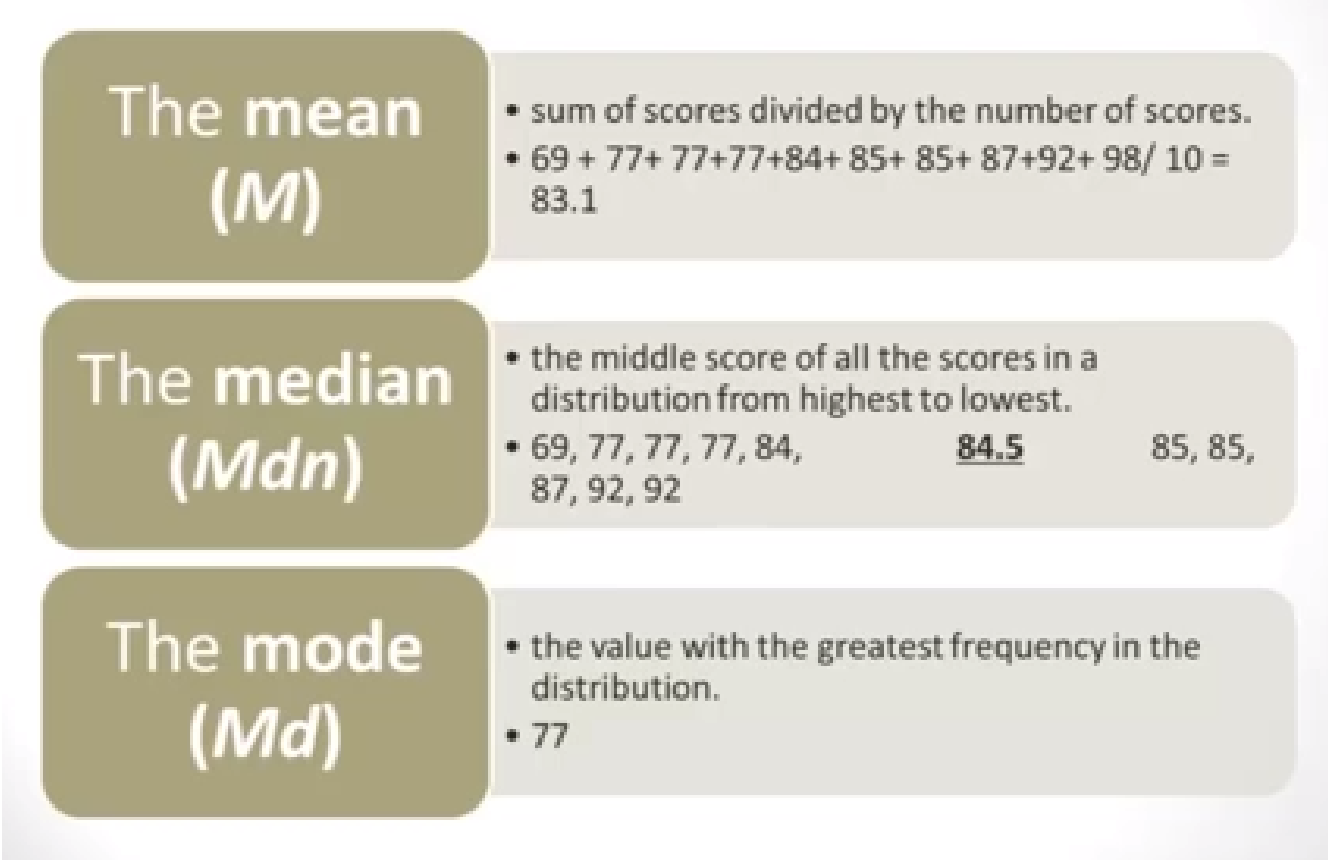
Measures of Frequency | **Measures of Central Tendency** | Measures of Spread | Measures of Position

Central tendency indicates whether the data values accumulate in the middle of distribution or toward the end.

The measures of central tendency are mean, median, and mode.

CENTRAL TENDENCY

- 69,77,77,77,84,85,85,87,92,98



STATISTICAL MEASURES

Measures of Frequency | Measures of Central Tendency | **Measures of Spread** | Measures of Position

Spread describes how similar or varied the set of observed values are for a particular variable.

The measures of spread are standard deviation, variance, and quartiles.

The measures of spread are also called measures of dispersion.

DISPERSION

- Range
  - the distance between the minimum and the maximum
- Standand deviation
  - Square root of the variance
    - Variance
      - The average of the squared differences from the mean
      - Calculate the mean
      - For each value, calculate the difference from the mean, and square each of the number
      - Calculate the mean of the squared differences
  - It measures the dispersion of scores around the mean
  - How spread out are the values from what is "normal" or average
  - The larger the SD means the larger the spread of values around the mean

STATISTICAL MEASURES

Measures of Frequency | Measures of Central Tendency | Measures of Spread | **Measures of Position**

Position identifies the exact location of a particular data value in the given data set.

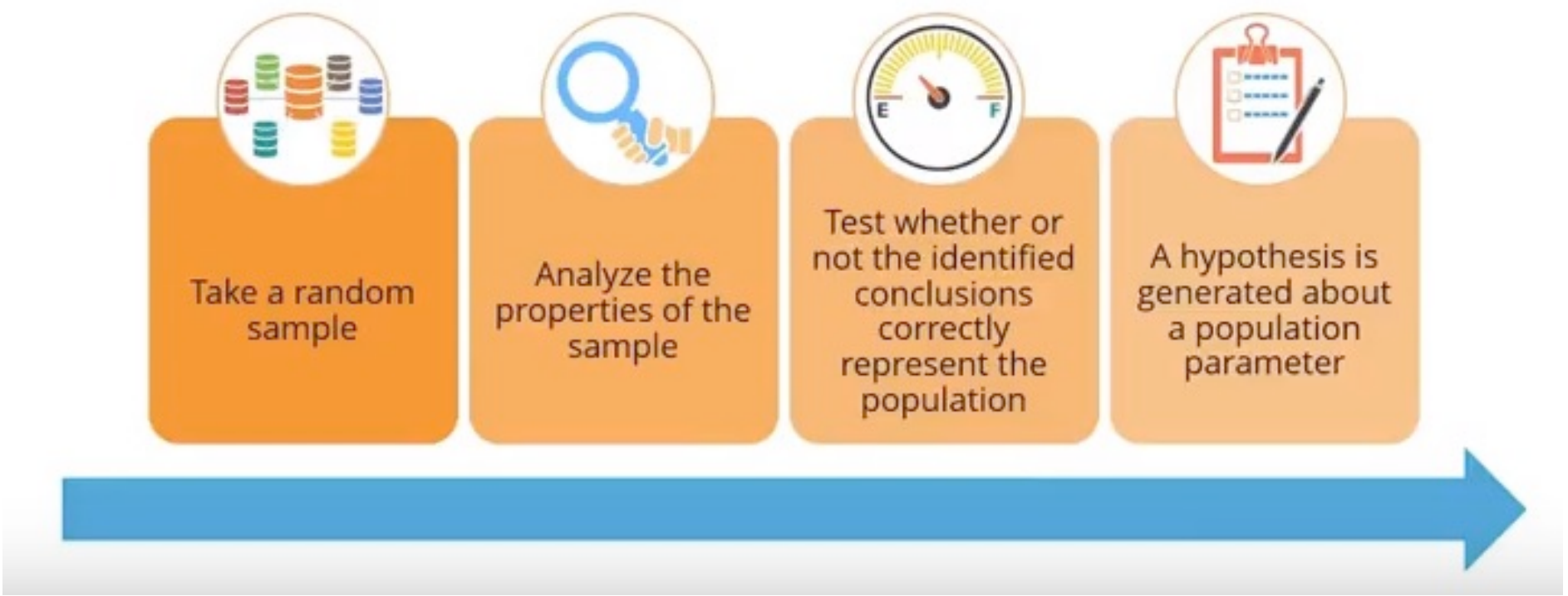
The measures of position are percentiles, quartiles, and standard scores.

MEASURES OF POSITION

- Z-score (Standard score)
  - The precise location of each X value within a distribution in relation to the mean
  - Gives the number of SD that a score lies
  - Positive sign indicates above or less than the mean
    - 0 = M
    - Negative or less than 0 = less than M
    - Positive or greater than 0 = greater than M
- $Z = (X - M) / s$
- Note : the critical Z score values when using a 95% confidence level are -1.96 and +1.96 standard deviations

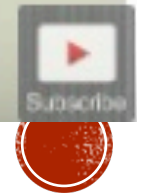
INFERENTIAL STATISTICS : HYPOTHESIS TESTING

- To determine if there is enough evidence in data sample to infer that a certain condition holds true for the entire population

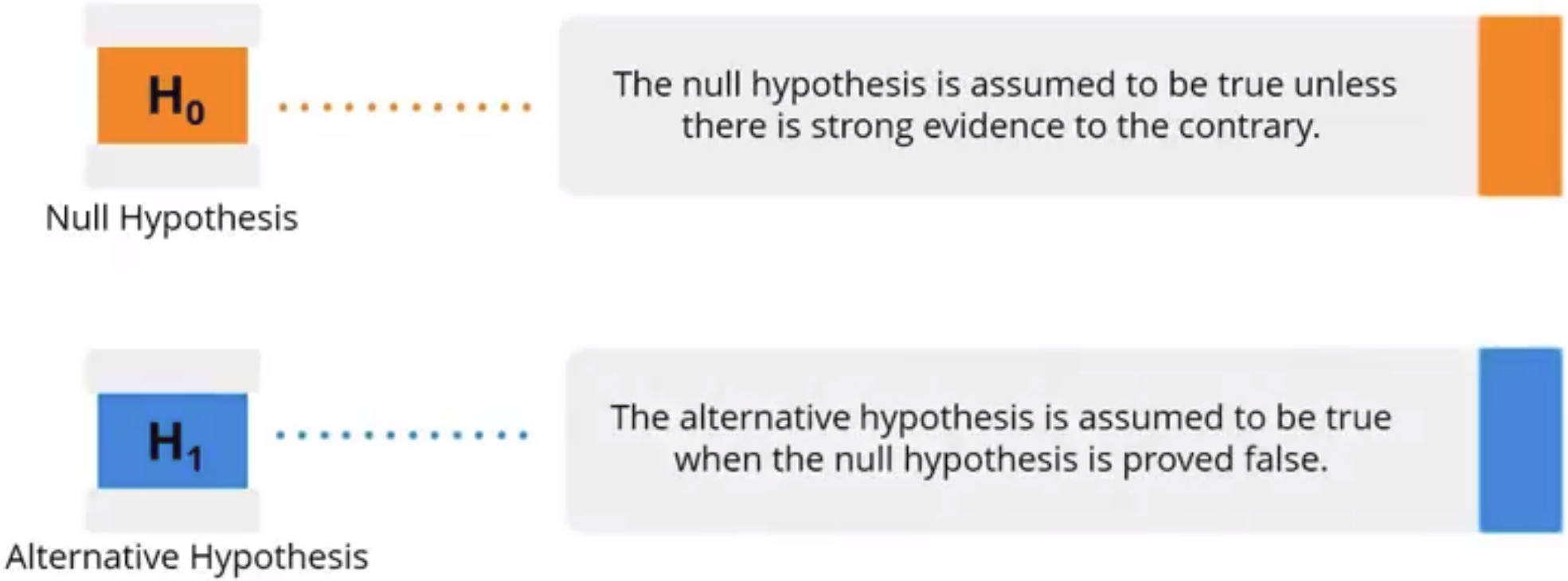


HYPOTHESIS TESTING PROCEDURE

1. Start with a well-developed, clear research problem or question
2. Establish hypotheses, both null and alternative
3. Determine appropriate statistical test and sampling distribution
4. Choose the Type I error rate
5. State the decision rule
6. Gather sample data
7. Calculate test statistics
8. State statistical conclusion
9. Make decision or inference based on conclusion



HYPOTHESIS TESTING



HYPOTHESIS EXAMPLES

- Null vs alternative

- Null : No variation exists between variables under study
- Alternative : Any hypothesis other than the null
- Alternative / Research hypothesis – a relationship or difference exists between variables under study

- The wages of men and women are equal (Null)
- The wages between men and women are not equal

- The medicine is safe (Null)
- If proven that the medicine is unsafe, then the null hypothesis is rejected

POPULATION VS SAMPLE

- Population (N) : ALL the members of a group that a researcher plans to focus on. Can be very large – cannot collect data from everyone in the population
- Sample (n) : a smaller group that represents a population. Want to generalize the result to a population from a smaller sample taken from that population. Assume n is randomly collected from population N that is normally distributed. Need to check these assumptions.

	Population	Sample
Mean	μ	M
Standard Deviation	σ	SD

STATISTICAL SIGNIFICANCE LEVEL

- Choosing alpha value :
  - 0.05 most often used (must be less than 0.05)
  - The chance of making type I error (REJECT the null H when it is ACTUALLY TRUE)
  - (eg. probably will make 1 error in 20 times)
  - We want to be at least 95% confident that when reject the null hypothesis, that it is the correct decision.

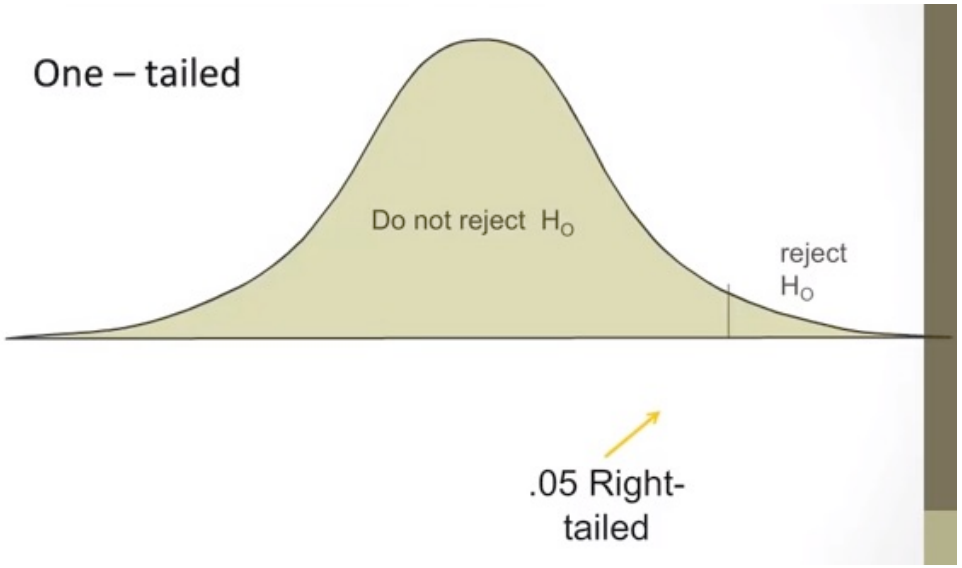
- Note : When doing 2-tailed test. Alpha 0.05 will be divided into two

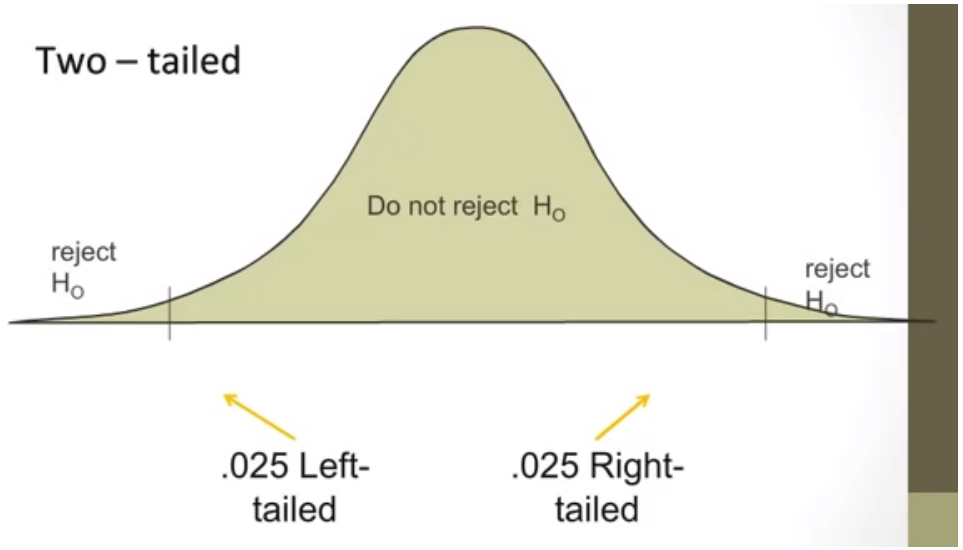
TYPE I VS TYPE II ERROR

- Type I error
  - When you REJECT the null hypothesis when the null hypothesis is TRUE
  - you reject H0 and you shouldn't

- Type II error
  - When you FAIL to reject when the null hypothesis is FALSE
  - you don't reject H0 and you should have

ONE-TAILED VS TWO-TAILED

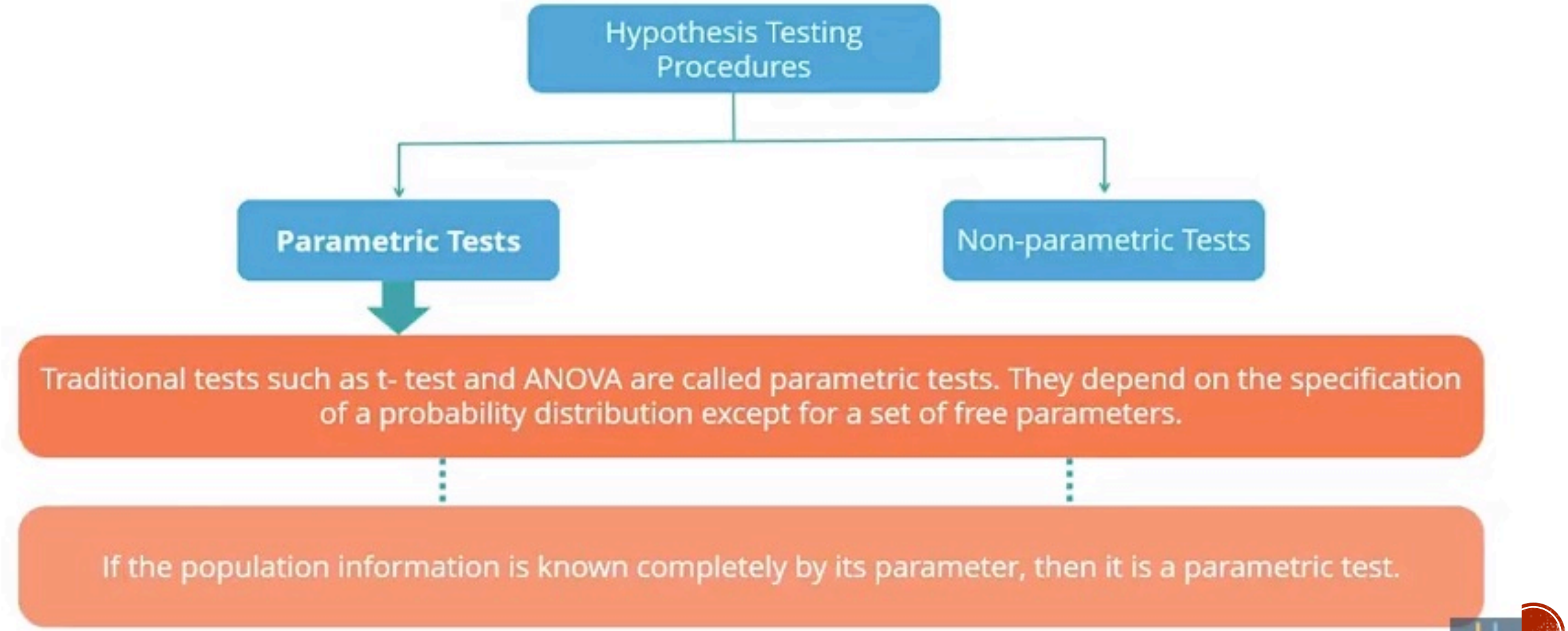




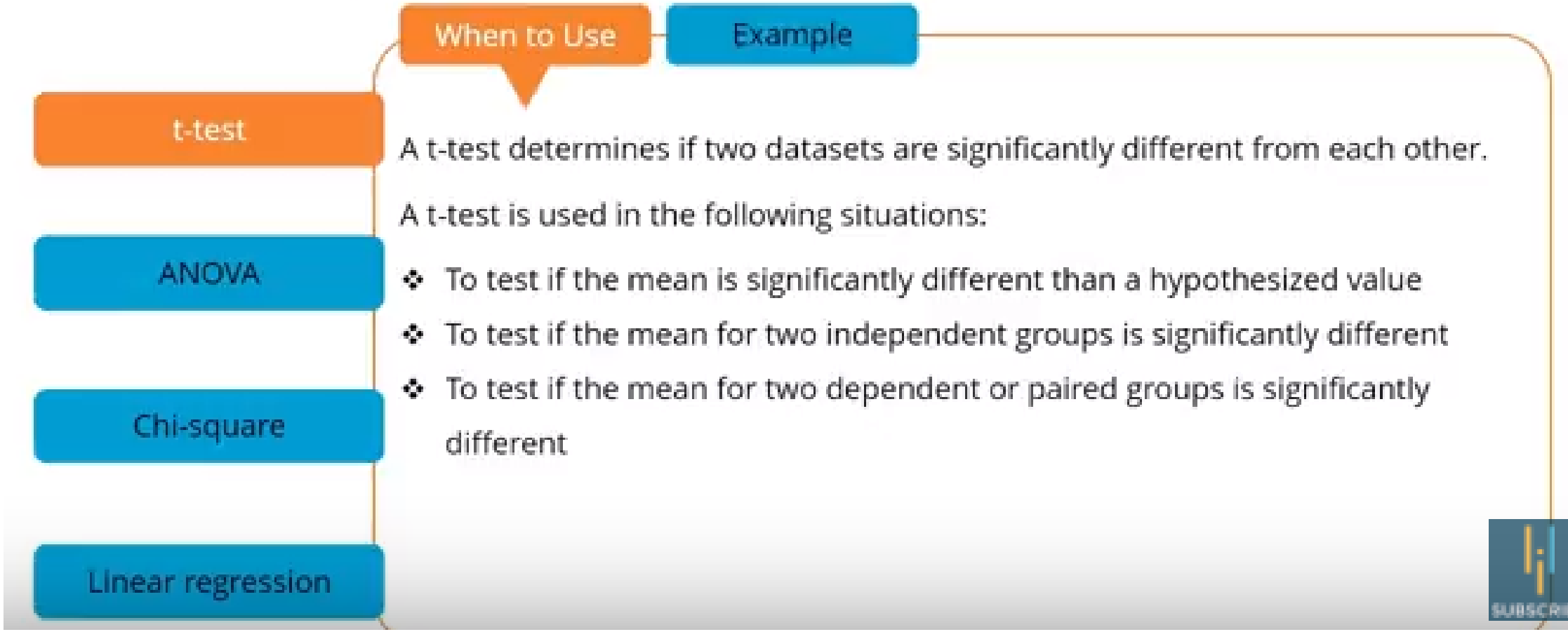
INTERPRET THE OUTPUT

- Alpha value is the probability of making an error, 5% is the standard (0.05)
- *p* value greater than the alpha value, thus cannot reject the hypothesis because NOT SATISTICALLY SIGNIFICANT
- *p* value less or equals the alpha value, the result is statistically significant THUS REJECT THE NULL
- *p* value CANNOT be zero

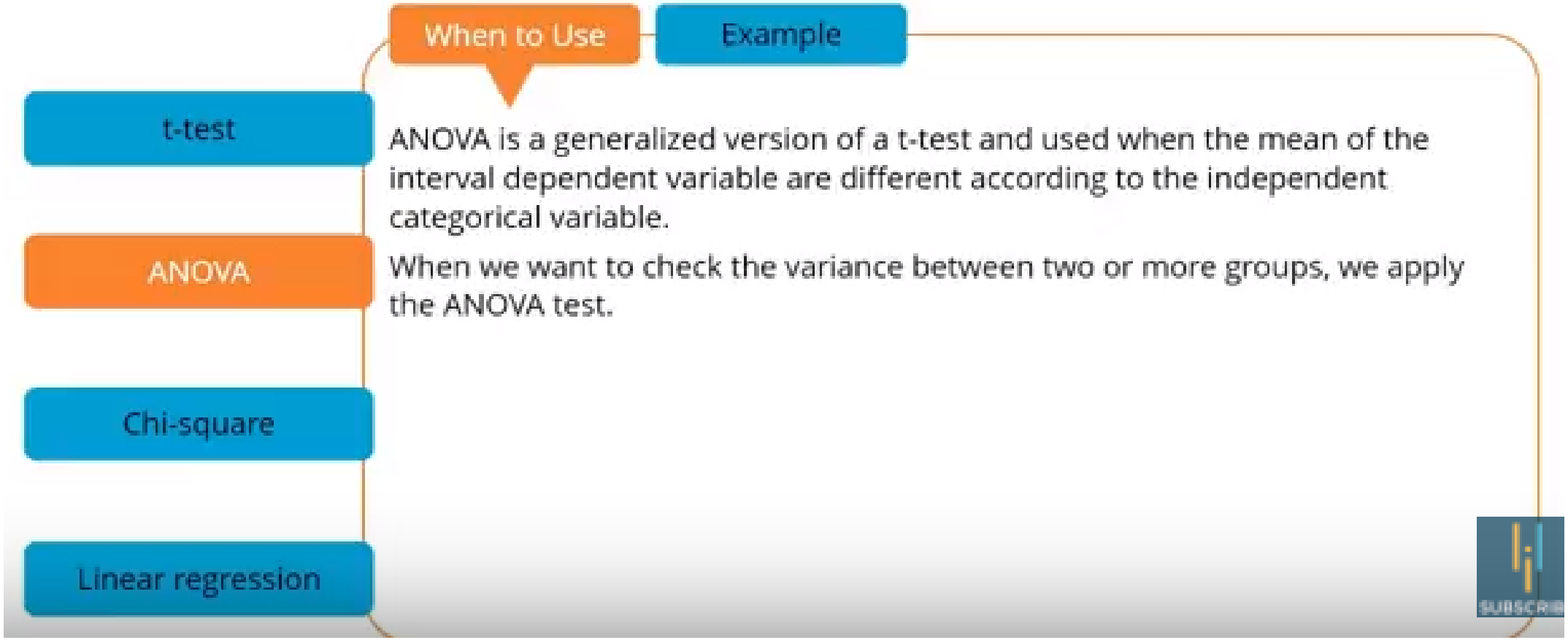
PARAMETRIC TEST



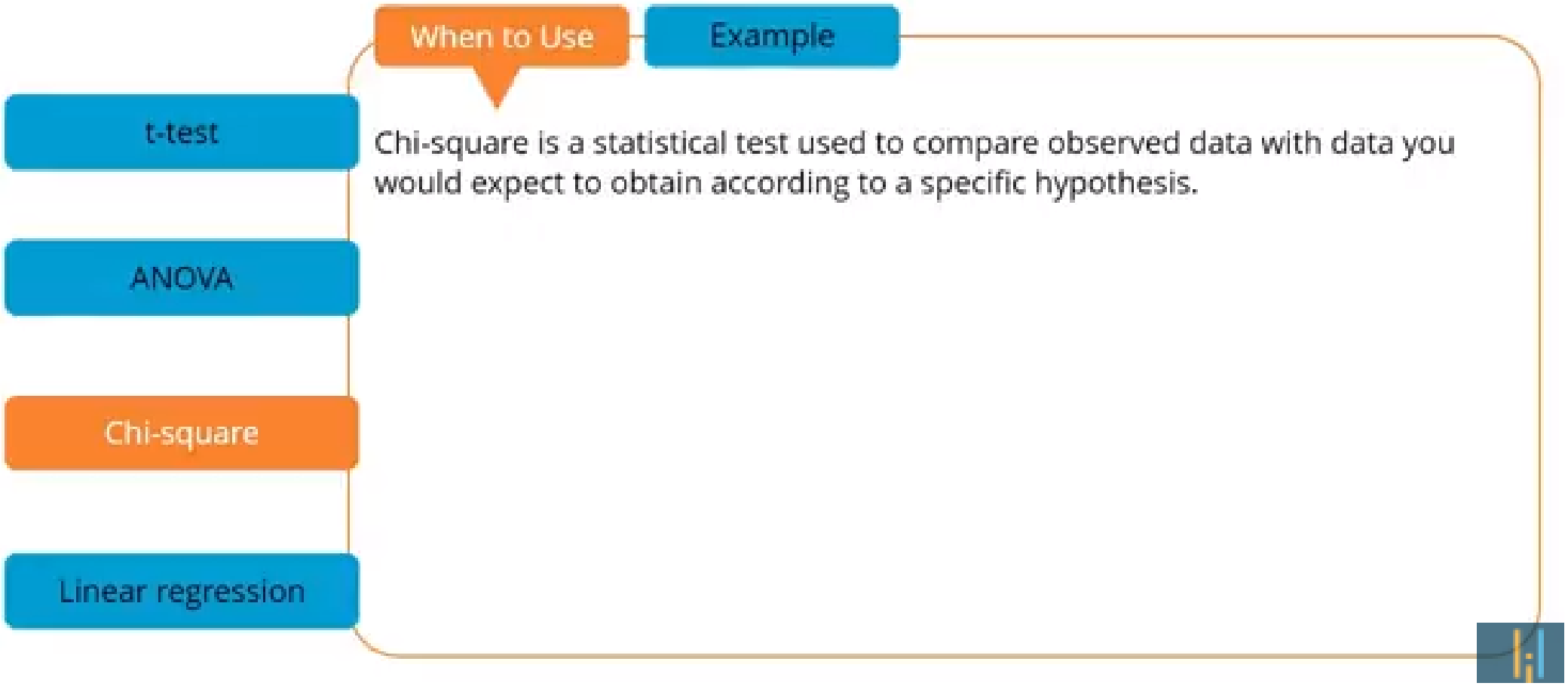
PARAMETRIC TESTS



PARAMETRIC TESTS

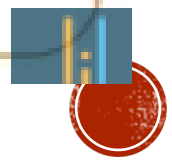


PARAMETRIC TESTS



PARAMETRIC TESTS

	When to Use	Example
t-test	There are two types of linear regression: simple linear regression and multiple linear regression.	
ANOVA	Simple linear regression is used to test how well a variable predicts another variable.	
Chi-square	Multiple linear regression is used to test how well multiple variables (Independent Variables) predict a variable of interest.	
Linear regression		



PARAMETRIC TESTS

When to Use | Example

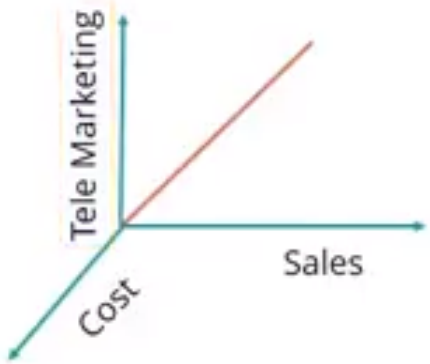
- t-test
- ANOVA
- Chi-square
- Linear regression

Simple Linear Regression

Finding a relationship between any two variables is called simple linear regression.

Multiple Linear Regression

Finding a relationship between any three variables is called multiple linear regression.

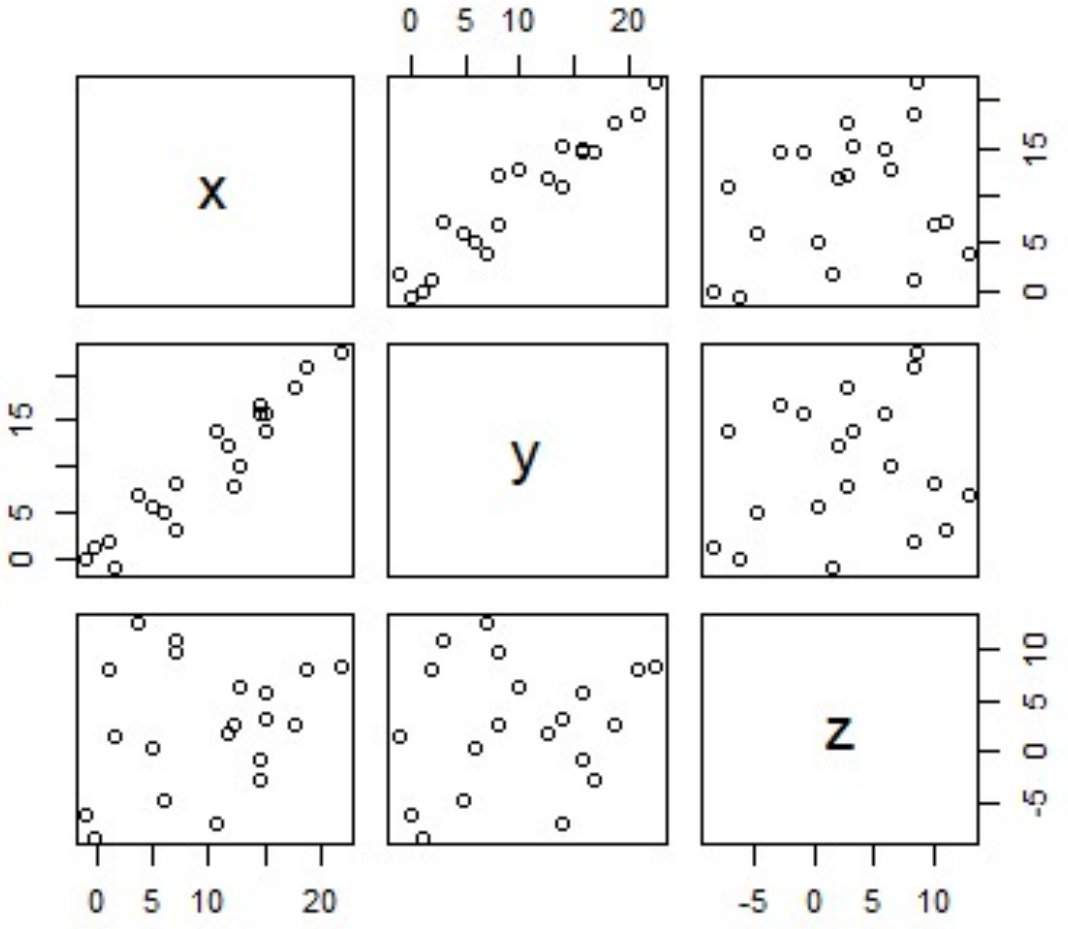


STATISTICAL ANALYSIS

- Correlation and linear regression each explore the relationship between two quantitative variables.
- Correlation determines if one variable varies systematically as another variable changes.
- Pearson, Kendall, and Spearman
- Linear regression specifies one variable as the independent variable and another as the dependent variable.

CORRELATION

```
cor(df,method="pearson")

          x         y          z
x 1.0000000 0.8736874 -0.2485967
y 0.8736874 1.0000000 -0.2376243
z -0.2485967 -0.2376243 1.0000000

cor(df[,1:3],method="spearman")

          x         y          z
x 1.0000000 0.8887218 -0.3323308
y 0.8887218 1.0000000 -0.2992481
z -0.3323308 -0.2992481 1.0000000
```



```
cor.test(df$x,df$y,method="pearson")
```

```
        Pearson's product-moment correlation

data:  df$x and df$y
t = 7.6194, df = 18, p-value = 4.872e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7029411 0.9492172
sample estimates:
      cor
0.8736874
```

```
cor.test(df$x,df$y,method="spearman")
```

```
        Spearman's rank correlation rho

data:  df$x and df$y
S = 148, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.8887218
```

REGRESSION MODEL

- Put variables into equation , explain causal relationship between them
- Y=mX+b
- Y = m*X + b

- **Y** is Dependent Variable
- **X** is Independent Variable
- **b** is intercept(mnemonic : 'b' means where the line begins)
- **m** is slope (mnemonic : 'm' means 'move')

REGRESSION MODEL

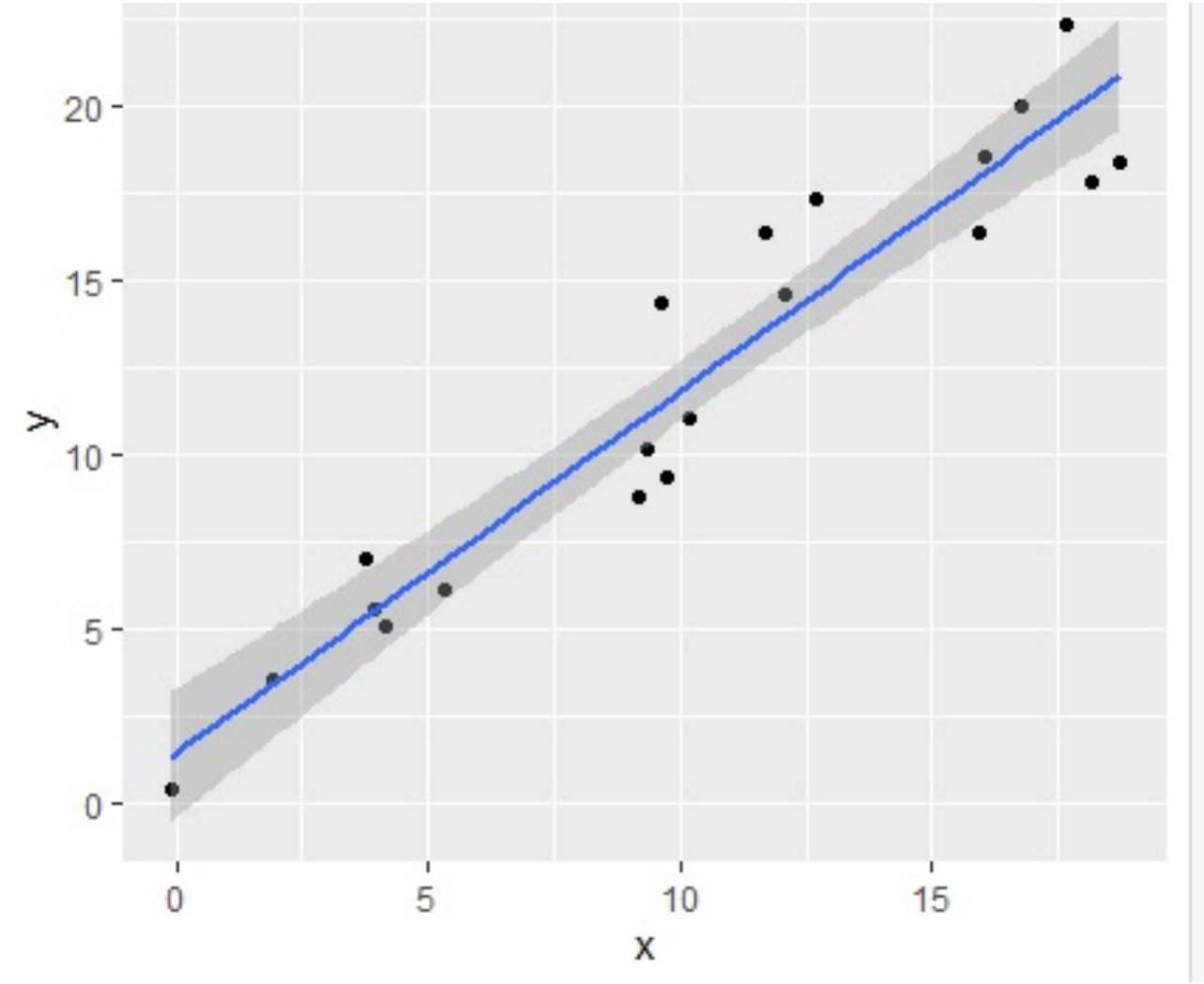
- Basic function to build linear model (linear regression) in R is lm()
- lm(mpg~cyl,data=mtcars)

  Or
- m1<- lm(mpg~cyl,data=mtcars)
- summary(m1) gives the summary of the regression model

- cor(df,method="pearson")

```
          x         y         z
x 1.0000000 0.8736874 -0.2485967
y 0.8736874 1.0000000 -0.2376243
z -0.2485967 -0.2376243 1.0000000
```

- ggplot(df, aes(x,y)) + geom_point()
  + geom_smooth(method = "lm")



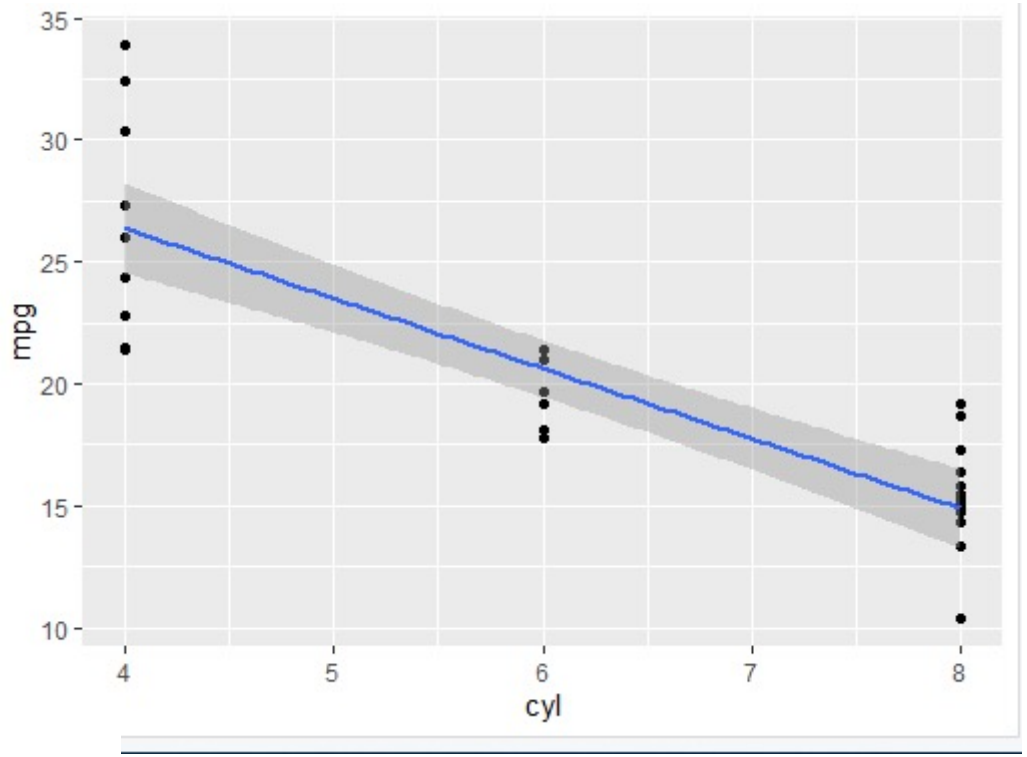
- lm(y~x,data=df)

```
Call:
lm(formula = y ~ x, data = df)

Coefficients:
(Intercept)            y
    -0.3332       0.8792
```

```
m2<-lm(y~x,data=df)
summary(m2)
```

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-2.64084 -1.02787  0.00115  1.04914  2.93892

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.33322    0.86062  -0.387    0.703
y            0.87924    0.06335  13.879 4.69e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.729 on 18 degrees of freedom
Multiple R-squared:  0.9145,     Adjusted R-squared:
0.9098
F-statistic: 192.6 on 1 and 18 DF,  p-value: 4.694e-11
```

```
> cor(mtcars, method="pearson")
             mpg        cyl       disp
mpg   1.0000000 -0.8521620 -0.8475514
cyl  -0.8521620  1.0000000  0.9020329
disp -0.8475514  0.9020329  1.0000000
hp   -0.7761684  0.8324475  0.7909486
drat  0.6811719 -0.6999381 -0.7102139
wt   -0.8676594  0.7824958  0.8879799
qsec  0.4186840 -0.5912421 -0.4336979
vs    0.6640389 -0.8108118 -0.7104159
am    0.5998324 -0.5226070 -0.5912270
gear  0.4802848 -0.4926866 -0.5555692
carb -0.5509251  0.5269883  0.3949769
```




```
> lm(mpg~cyl,data=mtcars)

Call:
lm(formula = mpg ~ cyl, data = mtcars)

Coefficients:
(Intercept)          cyl
     37.885       -2.876
```

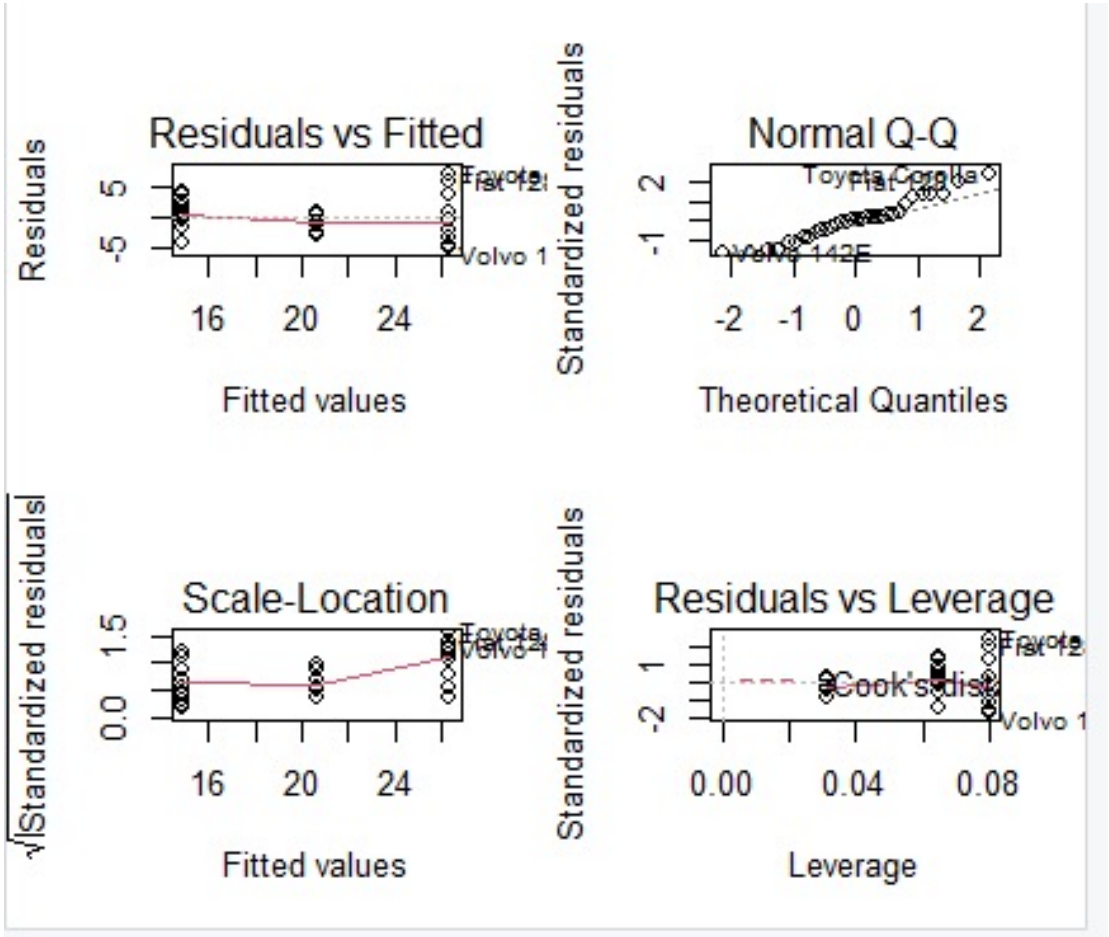
Additional Note :

```
m1<-lm(mpg~cyl,data=mtcars)
summary(m1)

par(mfrow=c(2,2))
plot(m1)
```

Use par() to put multiple plots alongside each other

But caution : your subsequent plots may be distorted, so you may have to reset back
Eg.

```
par(mfrow=c(1,1))
```

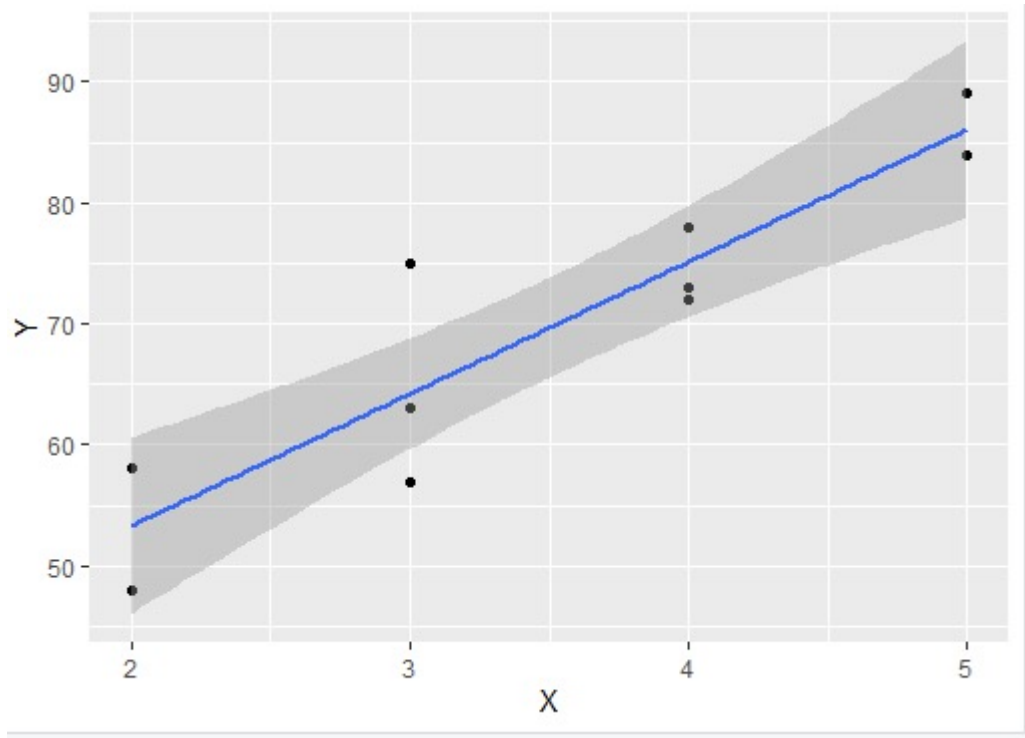


```
> anova(m1)
Analysis of Variance Table

Response: mpg
           Df Sum Sq Mean Sq F value    Pr(>F)
cyl         1 817.71  817.71  79.561 6.113e-10 ***
Residuals 30 308.33   10.28
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

PREDICTING WITH LINEAR REGRESSION

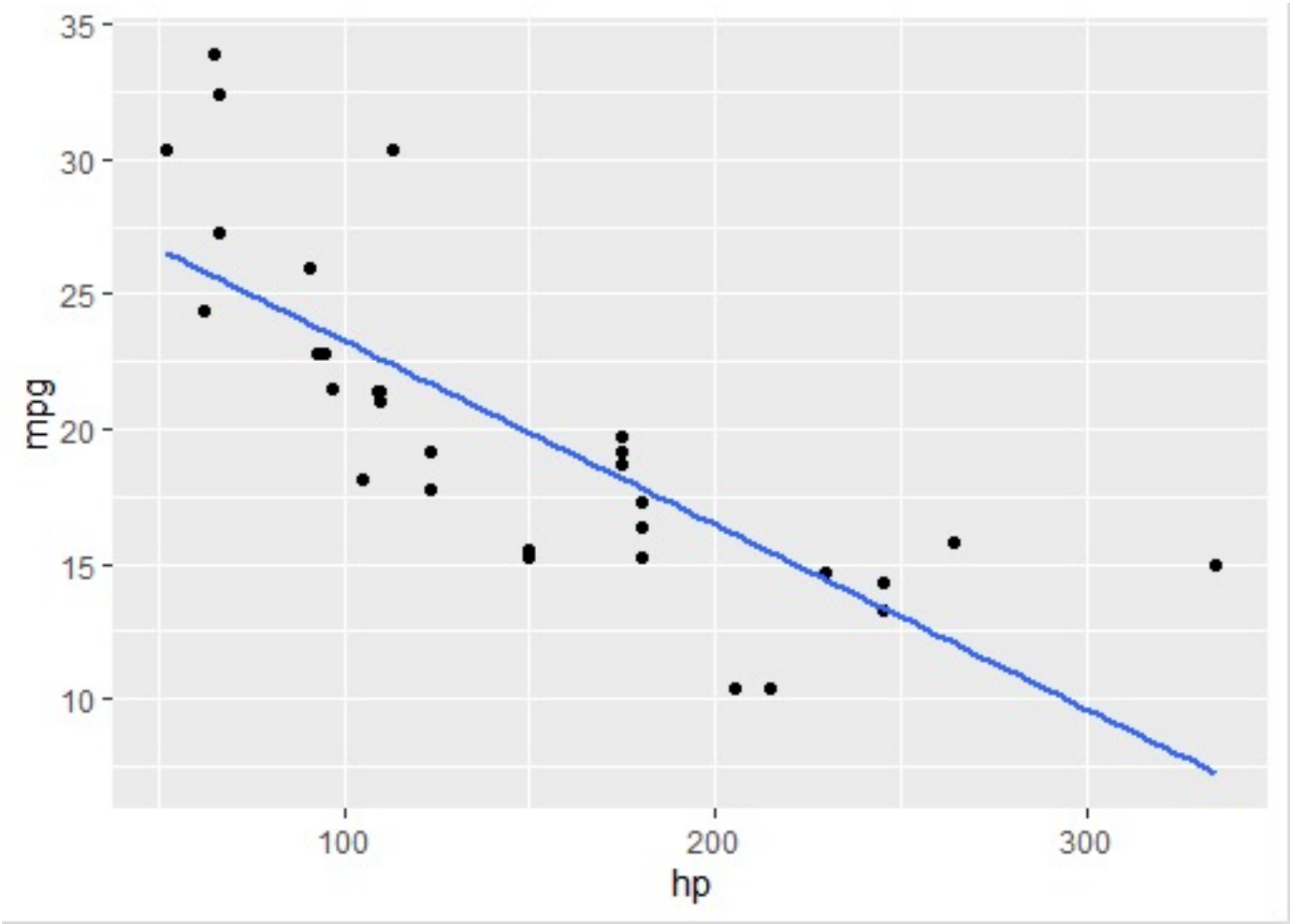
```
X = c(3,4,4,2,5,3,4,5,3,2)
Y= c(57,78,72,58,89,63,73,84,75,48)
DF<-data.frame(X,Y)
plot(DF)
ggplot(DF, aes(X,Y)) + geom_point() + geom_smooth(method = "lm")
cor(DF)
cor.test(DF$X, DF$Y)
F = lm(Y~X)
summary(F)

predict(F, newdata = data.frame(X=4.5), interval = "prediction")
```

```
> summary(F)

Call:
lm(formula = Y ~ X)

Residuals:
   Min     1Q Median     3Q    Max
-7.248 -2.902 -1.652  2.919 10.752

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   31.533      6.360   4.958 0.001110 **
X             10.905      1.744   6.253 0.000245 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.651 on 8 degrees of freedom
Multiple R-squared:  0.8301,    Adjusted R-squared:  0.8089
F-statistic: 39.09 on 1 and 8 DF,  p-value: 0.000245

> predict(F, newdata = data.frame(X=4.5), interval = "prediction")
       fit      lwr      upr
1 80.60476 66.35716 94.85236
```



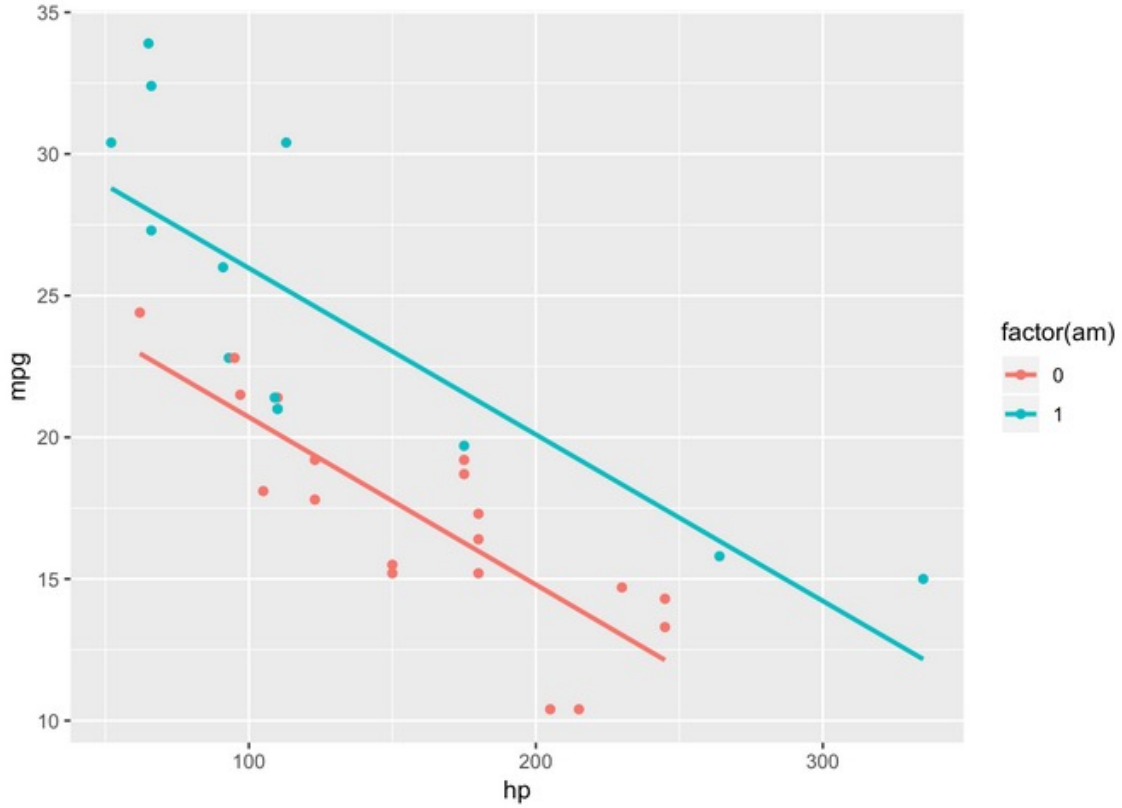
- #predict mpg based on hp

- to_correlate <- mtcars %>% dplyr::select(mpg, qsec, cyl, disp, hp)
- plot(to_correlate)
- ggplot(to_correlate, aes(y=mpg, x=hp)) + geom_point() + geom_smooth(method="lm", se=FALSE)
- ggplot(to_correlate, aes(y=mpg, x=hp)) + geom_jitter(width=0.1) + stat_smooth(method="lm", se=FALSE)
- cor(to_correlate)
- cor.test(mtcars$mpg, mtcars$hp)
- FC = lm(mpg~hp, data=mtcars)
- summary(FC)
- predict(FC, newdata = data.frame(hp=100), interval = "prediction")




```
> predict(FC, newdata = data.frame(hp=100), interval = "prediction")
       fit      lwr      upr
1 23.27603 15.2066 31.34547
```

MULTIPLE REGRESSION

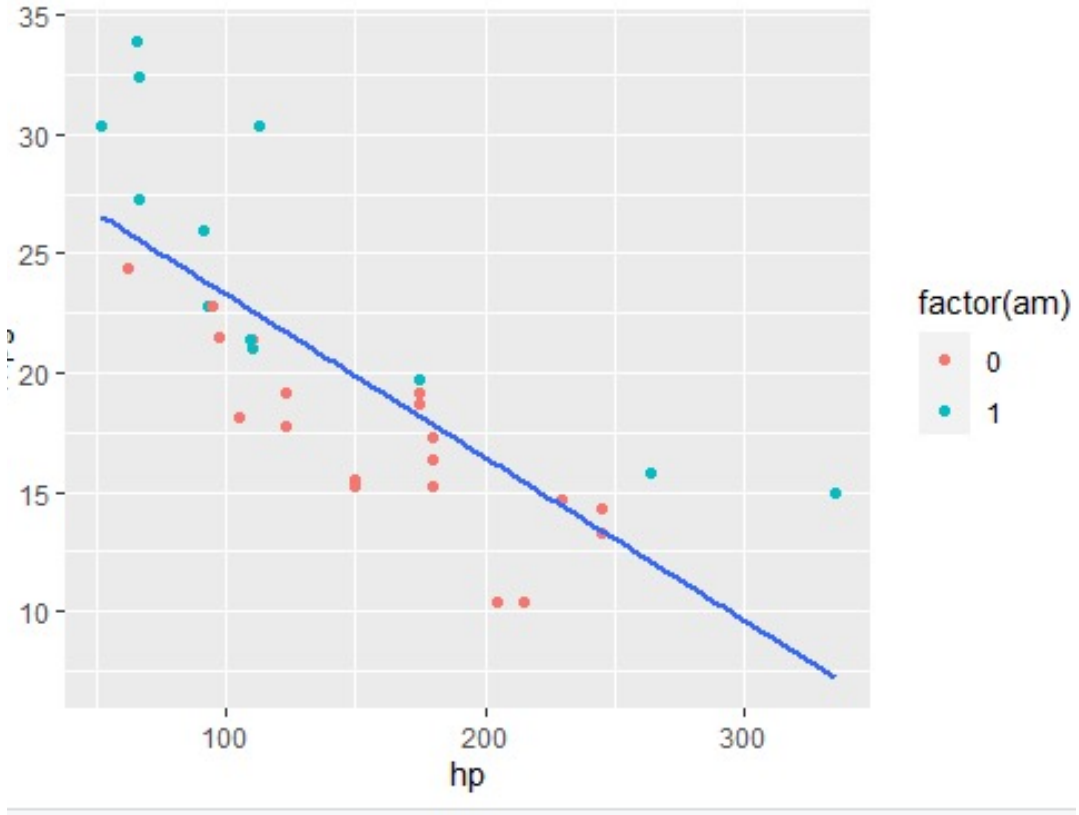


- For example
  - To examine the extent to which the gas mileage (mpg) is a function of

    both gross horsepower (hp) and transmission

    (am) where 0 is 'automatic' and 1 is 'manual')
- ggplot(mtcars, aes(x=hp, y=mpg, color=factor(am))) + geom_point() + stat_smooth(method=lm, se=FALSE)
- mpg_model <- lm(mpg ~ hp + am, mtcars)
- summary(mpg_model)

PREDICT MULTIPLE LINEAR MODEL



- ggplot(i_model, aes(x=hp, y=mpg)) + geom_point(aes(color= factor(am))) + geom_smooth(method="lm", se=FALSE)
- i_model <- lm(mpg ~ hp + am, mtcars)
- summary(i_model)
- prediction1 = predict(i_model, newdata = data.frame(hp = 110,am = 1), interval = "prediction")
- prediction1

```
> prediction1
       fit      lwr      upr
1 25.38434 19.20383 31.56485
> 
```

REFERENCES

- https://www.youtube.com/watch?v=Lv0xcdeXaGU
- https://www.youtube.com/watch?v=8Iklj-lf1fY
- https://www.youtube.com/watch?v=ZkEjYloGRIE
- https://www.youtube.com/watch?v=ACuy8Ab8CXA (writing H directional vs non directional)
- https://www.youtube.com/watch?v=prEM-vTfYKI (p value)
- https://www.youtube.com/watch?v=IC_z7EOu7sY
- https://www.youtube.com/watch?v=FiXCm3nGuyI
- https://www.youtube.com/watch?v=UaptUhOushw (choosing which test to use)

REFERENCES

- https://www.youtube.com/watch?v=NQWZefn41VY
- https://www.youtube.com/watch?v=kvmSAXhX9Hs (R one sample t test)
- https://www.youtube.com/watch?v=YsalXF5POtY (use z test or t test)
- https://www.youtube.com/watch?v=Kzqm8F9Le_4 (z vs t)