PYTHON PACKAGES OR LIBRARIES FOR DATA SCIENCE

Scientific Computing

Pandas — Data structures & tools
2D dataframes

NumPy — Arrays & Matrices

SciPy — Optimization and solving differential equations

zoom

PYTHON PACKAGES OR LIBRARIES FOR DATA SCIENCE

Data Visualization

Matplotlib — Plots, graphs and figures

Seaborn — heat maps, times series and other plots

zoom

PYTHON PACKAGES OR LIBRARIES FOR DATA SCIENCE

ML Algorithmic Development

Scikit-learn

Machine learning: regression, classification, clustering analysis and so on...

Statsmodels

Explore data, estimation of statistical models, and perform statistical analysis

MAIN OBJECTIVES

THIS WILL ENHANCE YOUR SKILLS:

1. Choosing a right statistical method
2. Do's and dont's of statistics
3. Reliable results
4. Paper revisions with proof of statistical test (Wh Qs)
5. Making Data Visualization
6. Interpreting results

Tests and their types

Parametric Tests

- More Reliable results
- First we have meet the assumptions

2	25
5	38
16	52
18	100
20	120

Not equal!

1	1
2	2
3	3
4	4
5	5

equal? based on the rankings

Non-Parametric tests

- Less reliable results
- Calculates the rank of data
- No need to m… the assump…

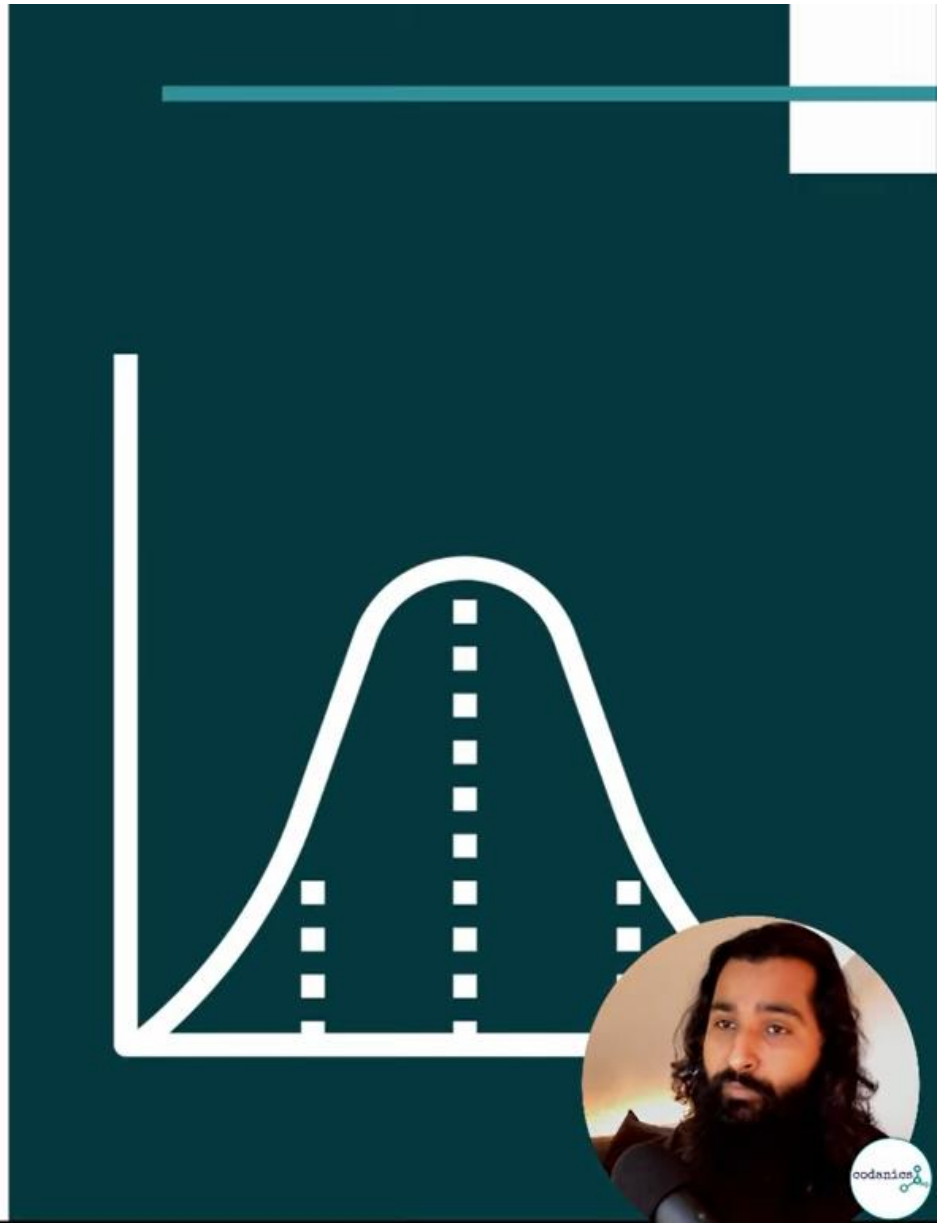
BEFORE STARTING

I repeat before starting the data analysis:

Step-1 Normality Test

Tests to be used:

1. **Shapiro-Wilk test**
   - Specific (Reliable)
2. **Kolmogorov-Smirnov Test**
   - General (Less reliable)



BEFORE STARTING

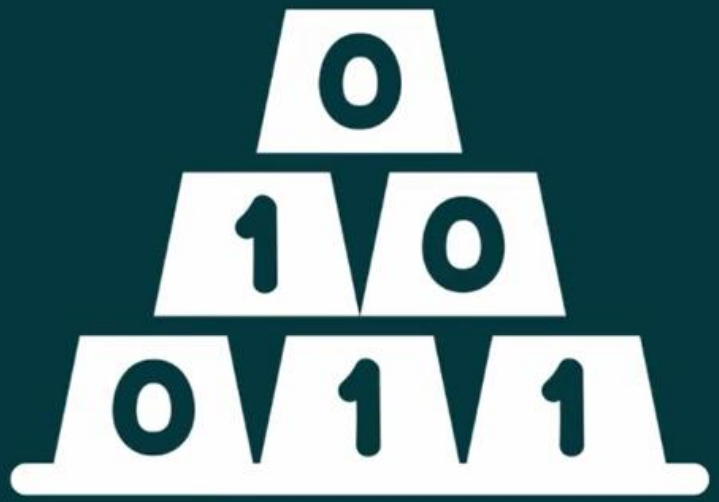
I repeat before starting the data analysis:

Step-2 Homogeneity Test

The variance of the variable in data are equal

Test to be used:

Levene's test

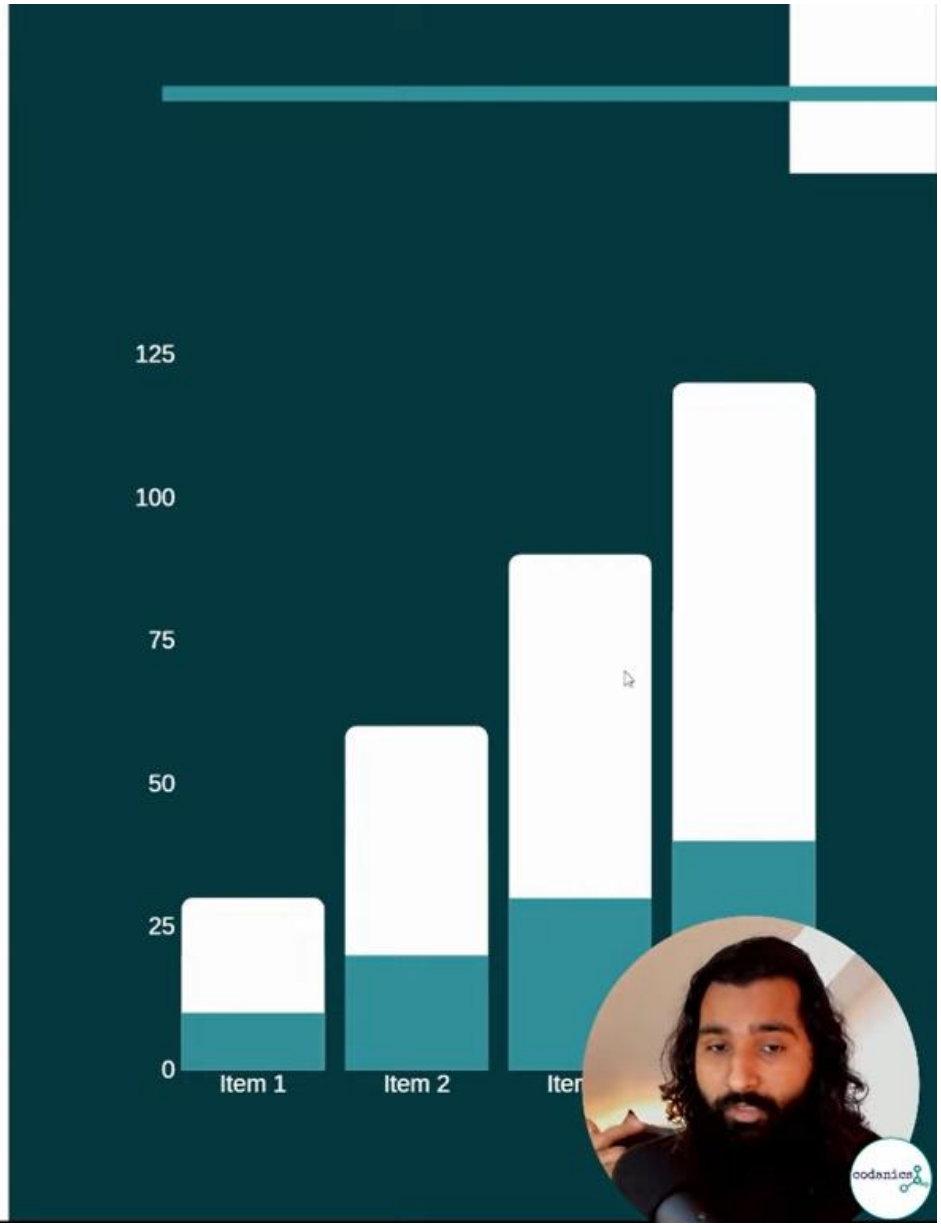


BEFORE STARTING

I repeat before starting the data analysis:

Step-3 Purpose

KNOW THE PURPOSE OF YOUR RESEARCH QUESTION



Two types of purposes

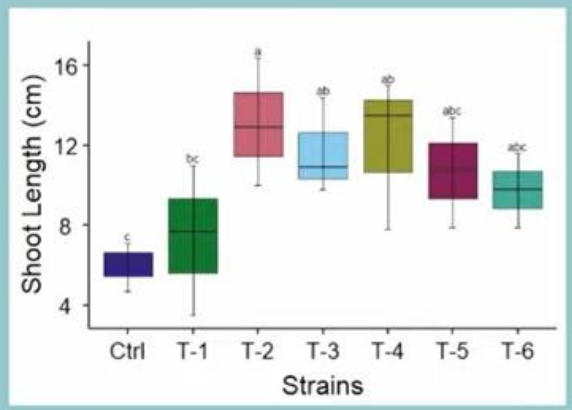
COMPARISON

DIFFERENCE

RELATIONSHIP

CONNECTION

Comparison



atleat two groups

EXAMPLES:

- Male vs Female
- Control group vs Treatments group
- Grouping individuals by color preference

Relationship

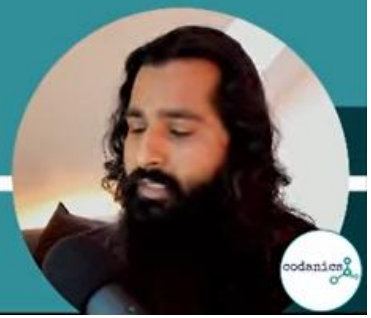
find a connection

EXAMPLES:

- Can food predict weight of a group of individuals
- Do fertilizer application increases crop growth?

We seek following here:

- **Connection**
- **Correlation**
- **Causation**
- **Prediction**



Step-4 Data Type

KNOW THE TYPE OF DATA YOU ARE WORKING WITH

Two types of Data

CATEGORICAL

Qualitative
No numerical meaning
Represented in texts
(e.g: character, factors)

CONTINUOUS

Quantitative
Numerical
Mostly represented in number
(e.g: Numerical variable, int and float)

Categorical

TRUE FALSE

YES OR NO

qualitative

EXAMPLES:

- Yes and No answers

(Have you ever been to Lahore?)

Which gene was expressed?

Do you like Mangoes? **"yes" or "No"**

Continuous

quantitative

EXAMPLES:

- Amount
- Number
- Age
- Plant Height
- Number of bacterial colonies
- Chlorophyll content
- Fertilizer Amount

Step-5 **Statistical Tests**

CHOOSE A STATISTICAL TEST FROM THREE MAIN FAMILIES

3 families of statiscal tests

1

Chi-Squared

Purpose: Comparison
Data: Categorical only
(Chi-Squared)

2

t-Test/ANOVA

Purpose: Comparison
Data: Categorical and Continuous
(t-Test)

3

Correlation

Purpose: Relationship
Data: Continuous only
(Correlation)

When and where to use?

1

Chi-Squared

Purpose: Comparison
Data: Categorical only
(Chi-Squared)

Types:

1. Chi-Squared test of homogeneity
2. Chi-squared test of independence

When to use?

- Nothing effects this,
- Can be used with any number of levels or groups

You must remember the purpose and data type

When and where to use?

2

t-Test/ANOVA

Purpose: Comparison
Data: Categorical and Continuous
(t-Test)

Types:

1. **One-sample t-Test** (for one sample group with a know mean)
2. **Two-sample t-Test**:
   - Un-paired t-Test (Two different groups)
   - Paired t-Test (Same group Twice)
3. **ANOVA** (Analysis of Variance) [3+ levels or groups are involved]
   - One-way ANOVA (Even one of group is significant you will get significant results, but doesn't tell you which one;)
   - Two-way ANOVA
   - Repeated measures of ANOVA (3+ paired groups, scale up of Paired t-Test)

When and where to use?

3

Correlation

Purpose: Relationship
Data: Continuous only (Correlation)

Types:

1. **Pearson's Correlation** (one-Independent and One-Dependent Variable)
2. **Regression** (one-Independent and One-Dependent Variable):

Correlation: Tells us how closely conencted two variables are?
"Is food a predictor of weight gain?"
Regression: Tells us a specific mathematical equation that describes the relationship.
(This helps us to find the data points not measured yet)
e.g: missing values can be predicted like this!

Important Things

Assumptions about your data

These tests trusts you that:

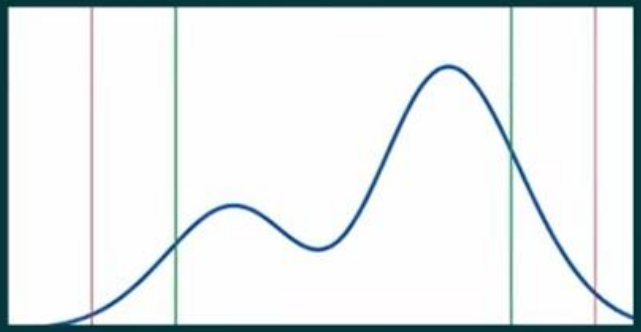
- Your data is Normally distributed
- or follow a Gaussian distribution

Non-reliable results

If you do not follow the assumptions and break the trusts of 3-test families, they will not happy with you!

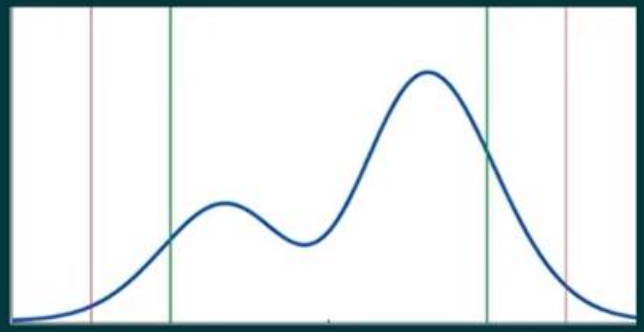
If,

Assumptions are not met!



If,

Assumptions are not met!



1. **Normalize your Data**
   a. Standardization
   b. Min-max scaling
   c. Log transformation
2. **Use Alternative Non-Parametric Tests**

Good News!

Non-parametric alternatives

1	2	3
Chi-Squared Purpose: Comparison Data: Categorical only (Chi-Squared)	**t-Test/ANOVA** 1. One-sample t-Test 2. Two-sample t-Test a. Unpaired t-Test b. Paired t-Test 3. ANOVA	**Correlation** Pearson's Correlation Regression
Chi-Squared	1. One-sample t-Test ◦ One-Sample Wilcoxon Signed rank test 2. Two-sample t-Test a. Unpaired t-Test (Mann Whitney's U-Test) b. Paired t-Test (Wilcoxon) 3. ANOVA (Kruskal-Wallis test)	**Pearson's Correlation** **(Spearman's Correlation)** **& (Kendall's Tau)** **Regression**

Important

Generalized and Simple

More to come for Advanced Level

Types of ANOVA

2

ANOVA's

Purpose: Comparison
Data: Categorical and Continuous
[3+ levels or groups are involved]

ANOVA (Analysis of Variance) [3+ levels or groups are involved]

1. One-way ANOVA (Even one of group is significant you will get significant results, but doesn't tell you which one;)
2. Two-way ANOVA (2 factors involved)
3. Repeated measures of ANOVA (3+ paired groups, scale up of Paired t-Test)

ANCOVA (Analysis of Co-variance)

- Compare the means of 3+ idenpendent groups which can not be tested by ANOVA because the variables are affected by co-variance (pre-test and post-Test of class)

MANOVA (Multi-variate analysis of Variance)

MANCOVA (Multi-variate analysis of Co-variance)

Some Other tests

Reliability tests

- Kunder-Richardson's Formula 20 and 21 (KR20/21)
- Cronbach's Alpha

Inter-rater Reliability tests

- Krippendorf's Alpha
  - (Categorical or continuous)
- Fleis's Kappa
  - (Only Categorical)

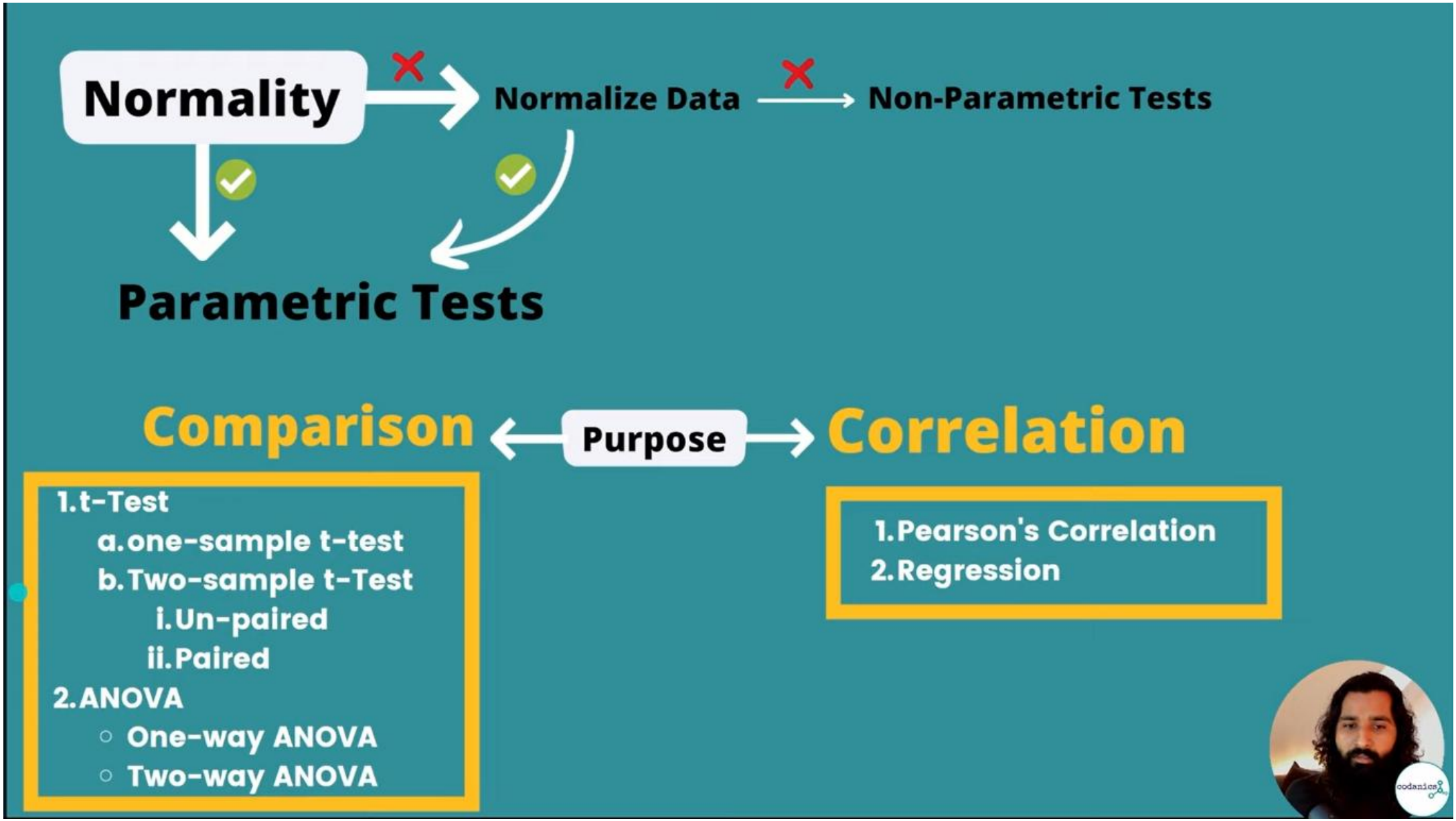
Validity tests

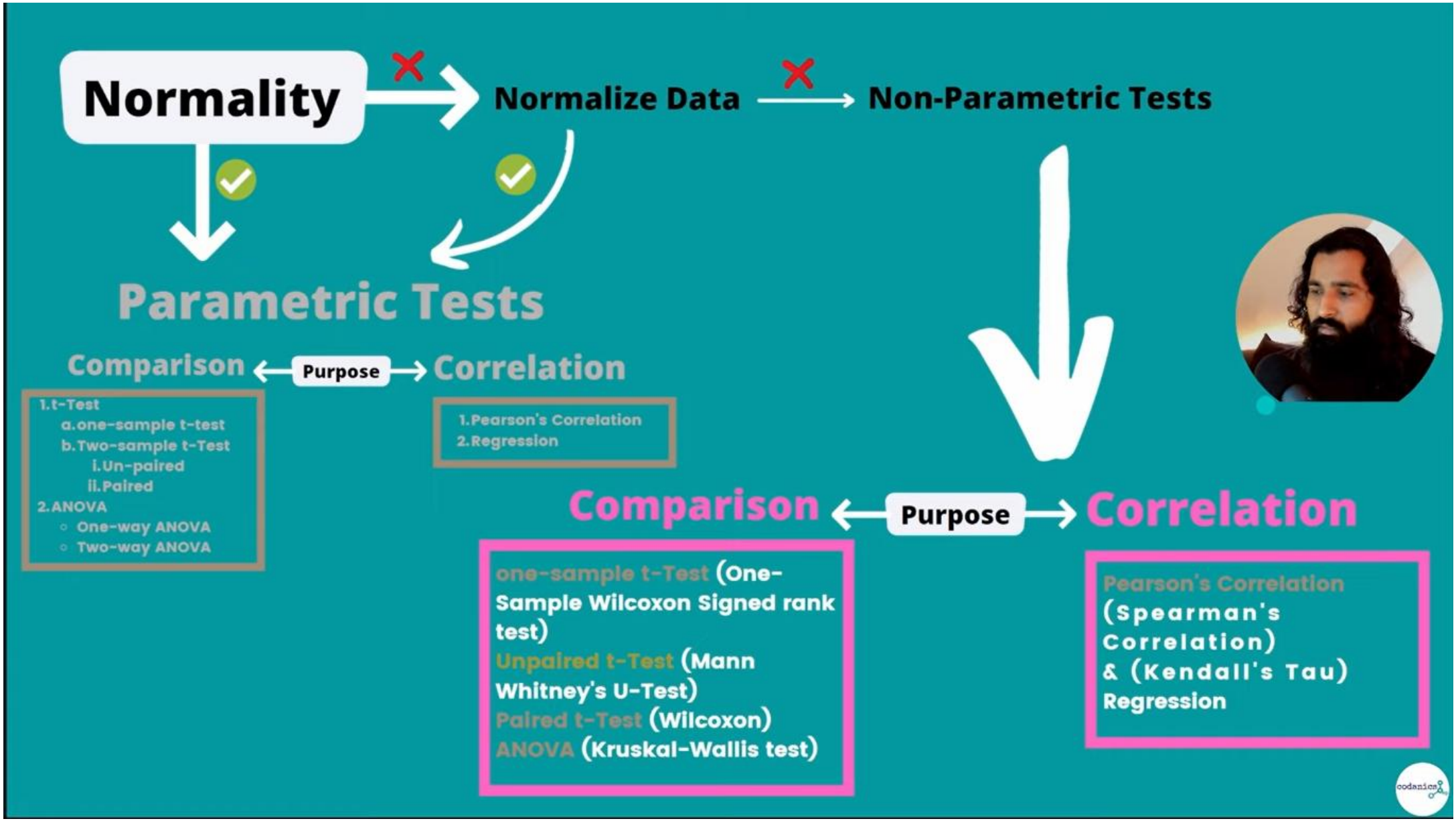
- Krippendorf's Alpha Test
- Fleis's Kappa Test

Sample size computation

How to make surehow many samples are valid?

- Cochran's Q Test
- Yamane's Test
- many others.....


Normality
Normalize Data
Non-Parametric Tests
Parametric Tests
Comparison
Purpose
Correlation
1. t-Test
a. one-sample t-test
b. Two-sample t-Test
i. Un-paired
ii. Paired
2. ANOVA
One-way ANOVA
Two-way ANOVA
1. Pearson's Correlation
2. Regression


Normality
Normalize Data
Non-Parametric Tests
Parametric Tests
Comparison
Purpose
Correlation
1.t-Test
a.one-sample t-test
b.Two-sample t-Test
i.Un-paired
ii.Paired
2.ANOVA
One-way ANOVA
Two-way ANOVA
1.Pearson's Correlation
2.Regression
Comparison
Purpose
Correlation
one-sample t-Test (One-Sample Wilcoxon Signed rank test)
Unpaired t-Test (Mann Whitney's U-Test)
Paired t-Test (Wilcoxon)
ANOVA (Kruskal-Wallis test)
Pearson's Correlation (Spearman's Correlation) & (Kendall's Tau)
Regression