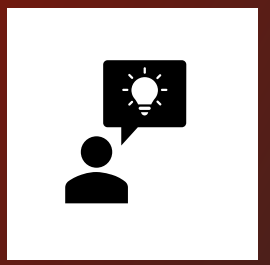
STATISTICAL TEST

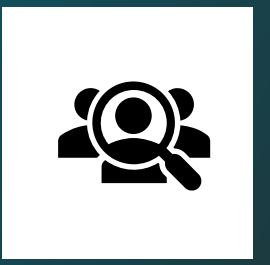
Assoc. Prof. Dr. Nurulhuda Firdaus Bt Mohd Azmi
huda@utm.my

SECOND SKILL: SPSS for Statistical Test



One Sample Statistical Test: Parametric and Non-Parametric

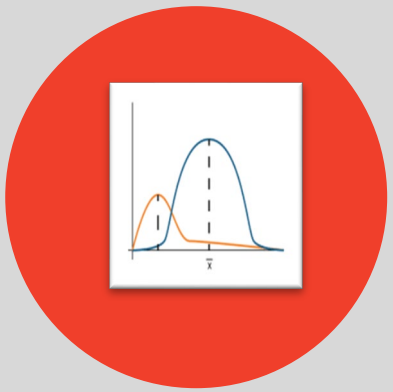
Independent Group Statistical Test: Parametric and Non-Parametric



Dependent (Paired) Group Statistical Test: Parametric and Non-Parametric

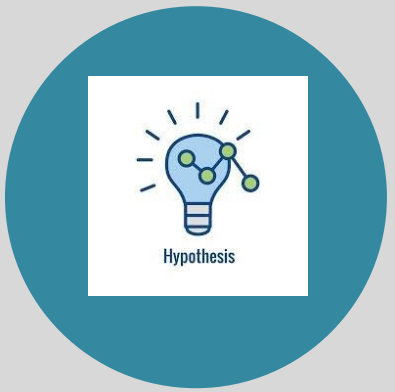
First Skill | Second Skill | Third Skill

STATISTICAL TEST

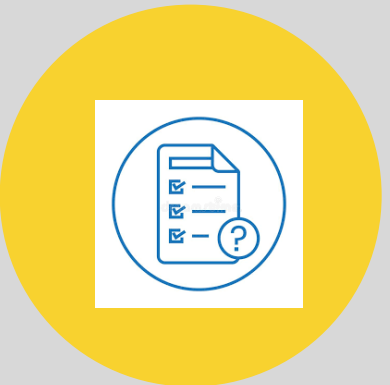


PARAMETRIC

NON-PARAMETRIC

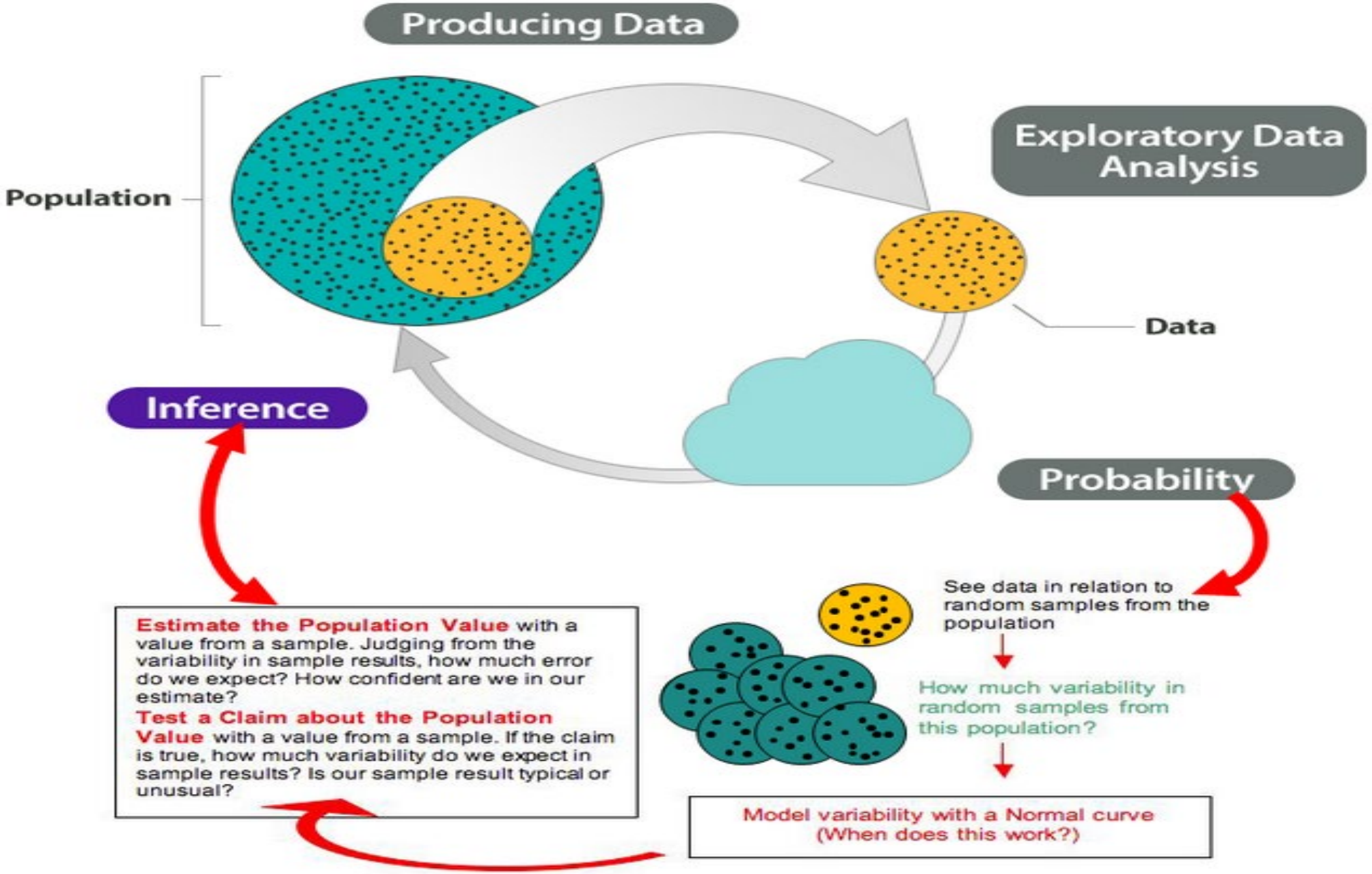


HYPOTHESIS TESTING
DECISION RULE

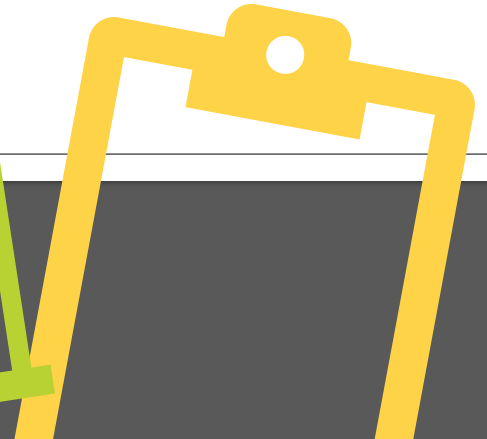


NORMALITY TEST

PARAMETRIC & NON-PARAMETRIC TEST


Producing Data
Population
Exploratory Data Analysis
Data
Inference
Probability
Estimate the Population Value with a value from a sample. Judging from the variability in sample results, how much error do we expect? How confident are we in our estimate?
Test a Claim about the Population Value with a value from a sample. If the claim is true, how much variability do we expect in sample results? Is our sample result typical or unusual?
See data in relation to random samples from the population
How much variability in random samples from this population?
Model variability with a Normal curve (When does this work?)

HYPOTHESIS TEST



HYPOTHESIS TEST

- Hypothesis test is performed regularly in many industries.
  - Example: Companies in the pharmaceutical industry must perform many hypothesis tests on new drug products before they are deemed to be safe and effective by the federal Food and Drug Administration (FDA).
- Describe a relationship among samples
  - Does the test samples are statistically difference or not
- Identify the significance of a variable
  - Does variable x is significance to be included in the prediction model?

HYPOTHESIS NULL & HYPOTHESIS ALTERNATIVE

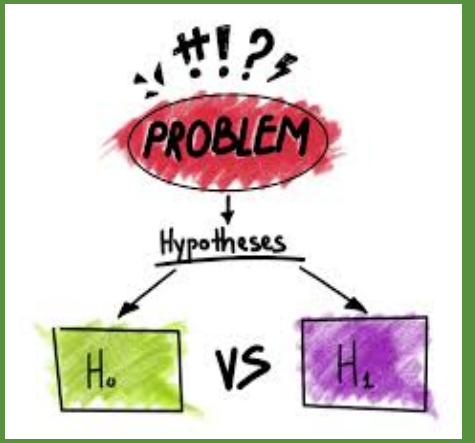
NULL HYPOTHESIS

Denoted as H_o

- A statement of no change, no effect or no difference and is assumed true until evidence indicates otherwise.
- Always contain a statement of equality

Example:

- The population mean time to answer customer emails was 4 hours last year.
- The mean height for women is the same as the mean height for men
- At a restaurant, the proportion of orders filled correctly for drive-through customers is the same as the proportion of orders filled correctly for sit-down customers.



ALTERNATIVE HYPOTHESIS

Denoted as H_1 or H_A

- Always contain a statement of non equality (either not equal or less than or greater than)
- Which would be paired with the null hypothesis

Example:

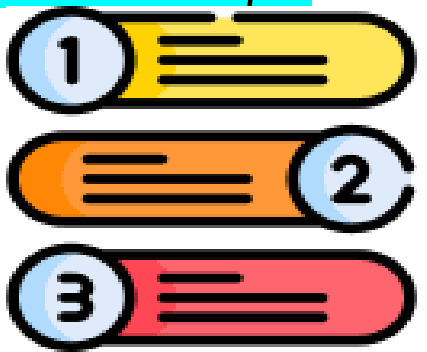
- The population mean for the time to answer customer complaints was more than 4 hours last year.
- The mean height for women is lower than the mean height for men
- The proportion of food orders filled correctly for drive through customers is not the same as the proportion of food orders filled correctly for sit-down customers

TWO SIDED: two tailed test	ONE SIDED (LEFT): one tailed test	ONE SIDED (RIGHT): one tailed test
H_o: parameter = some value	H_o: parameter ≥ some value	H_o: parameter ≤ some value
H_1: parameter ≠ some value	H_1: parameter < some value	H_1: parameter > some value
Example: H_o: μ = 20kg H_1: μ ≠ 20kg	Example: H_o: μ ≥ 20kg H_1: μ < 20kg	Example: H_o: μ ≤ 20kg H_1: μ > 20kg

BASIC PAIRS IN HYPOTHESIS

STEPS IN PERFORMING STATISTICAL HYPOTHESIS TEST

1. State the null hypothesis, H_0, and the alternative hypothesis, H_1.
2. Evaluate the risks of making type I and II errors, and choose the level of significance, α.
3. Collect the sample size as appropriate.
4. Calculate the *p*-value based on the test statistic and compare the *p*-value to α.
5. Decide the proper statistical inference. Reject the null hypothesis if the *p*-value is less than α. Do not reject the null hypothesis if the *p*-value is greater than or equal to α.



WHAT IS P-VALUE?

P-value Concept:

The probability of computing a test statistic equal to or more extreme than the sample results, given that the null hypothesis H_0 is true.

Decision Rule for p-values:

- If the *p*-value is greater than or equal to α, do not reject the null hypothesis.
- If the *p*-value is less than α, reject the null hypothesis.

To avoid confusion, remember this:

"The smallest the *p*-value, then H_0 must go."

FOUR OUTCOME FROM HYPOTHESIS TESTING

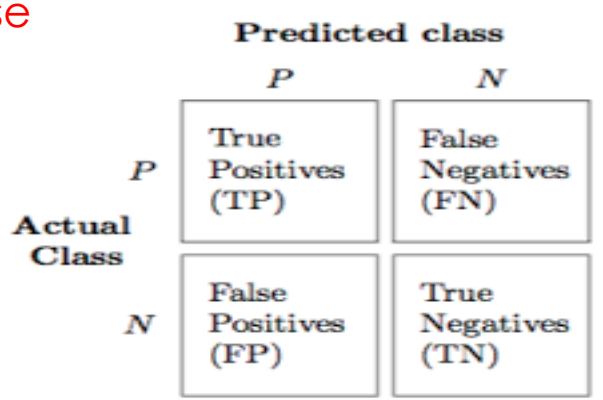
1. Reject the null hypothesis when the alternative hypothesis is true. This decision would be **correct**.
2. Do not reject the null hypothesis when the null hypothesis is true. This decision would be **correct**.
3. Reject the null hypothesis when the null hypothesis is true. This decision would be **incorrect**. This type of error is called a **Type I error**.
4. Do not reject the null hypothesis when the alternative hypothesis is true. This decision would be **incorrect**. This type of error is called a **Type II error**.

Risks and Decisions in Hypothesis Testing

		Actual Situation	
		H_0 True	**H_0 False**
Statistical Decision	Do not reject H_0	**Correct decision** Confidence = 1 – α	**Type II error** *P*(Type II error) = β
	Reject H_0	**Type I error** *P*(Type I error) = α	**Correct decision** Power = 1 – β

Col 1, row 1: you do not reject the H null when H null is true
Col 2, row 1: you do not reject the H null when H null is false (Type II:FN)
Col 1, row 2: you reject the H null when H null is true (Type I:FP)
Col 2, row 2: you reject the H null when H null is false

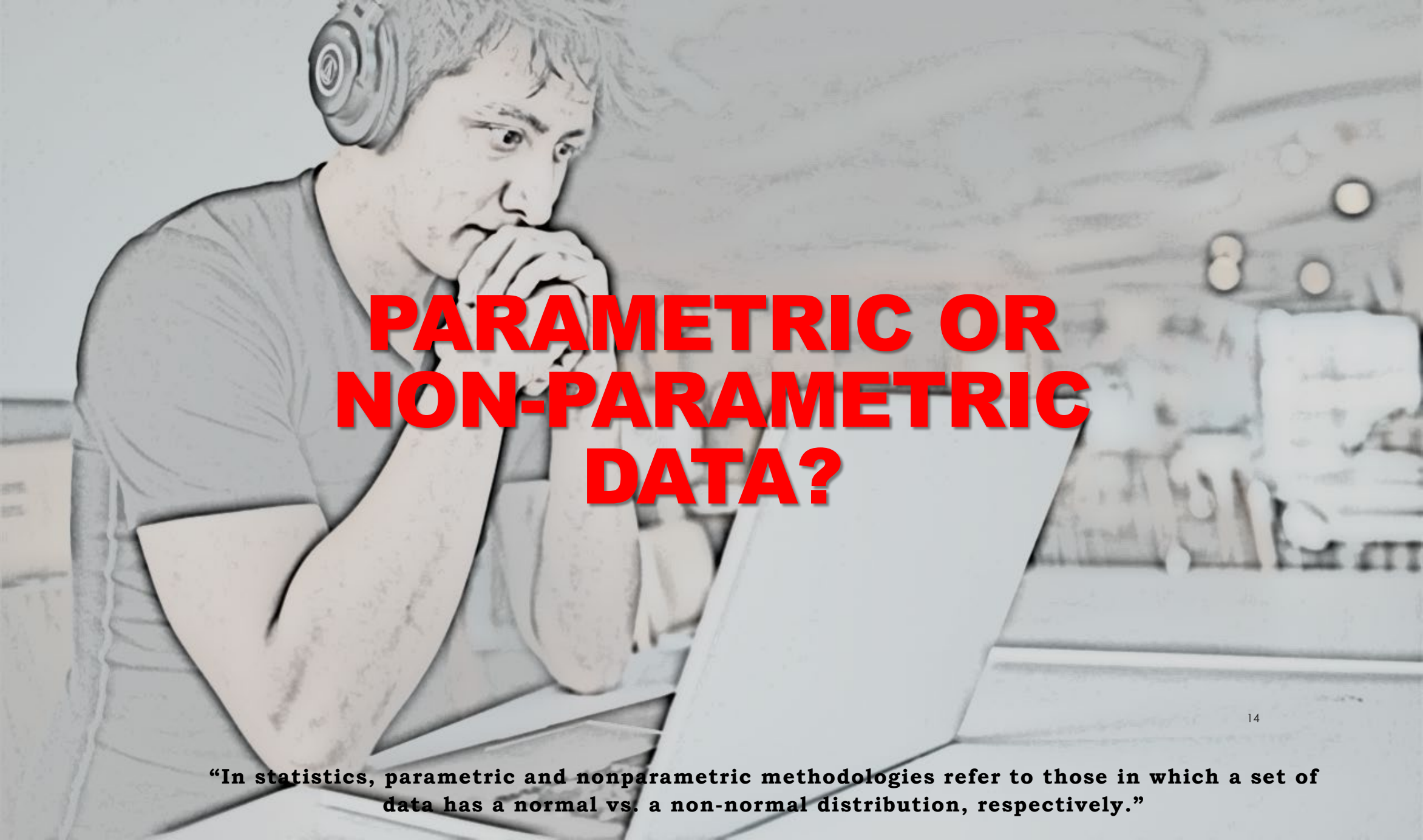
Confusion matrix in data mining



CAUTION!!!

- Cannot state 100% certainty that the statement is TRUE because sample data is used to test hypothesis.

- Can only determine whether the sample data support the (hypothesis) statement or does not support the (hypothesis) statement.



PARAMETRIC OR NON-PARAMETRIC DATA?

“In statistics, parametric and nonparametric methodologies refer to those in which a set of data has a normal vs. a non-normal distribution, respectively.”

WHY NEED TO CHECK PARAMETRIC AND NON-PARAMETRIC DATA?

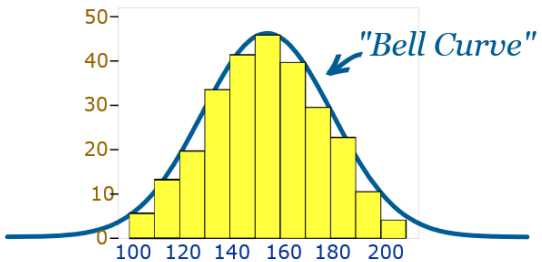
Statistical Assumption ➔ Normality Assumption

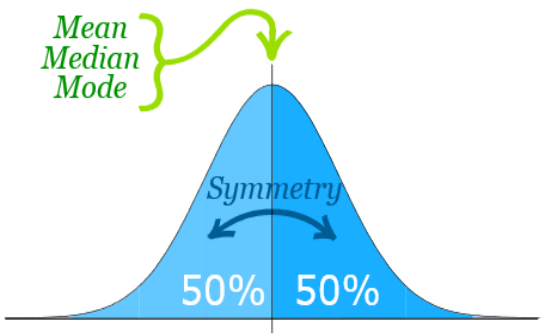
Normality Assumption:

- Data should follow **normal distribution**.
- For data that **follow** normal distribution, it is parametric data (which need to use parametric method).
- For data that is **not follow** normal distribution, it is non- parametric data (which need to use non-parametric method).

NORMAL DISTRIBUTION

- Also known as the Gaussian distribution and the bell curve.
- Data tends to be around a central value with no bias left or right.





We say the data is "normally distributed":

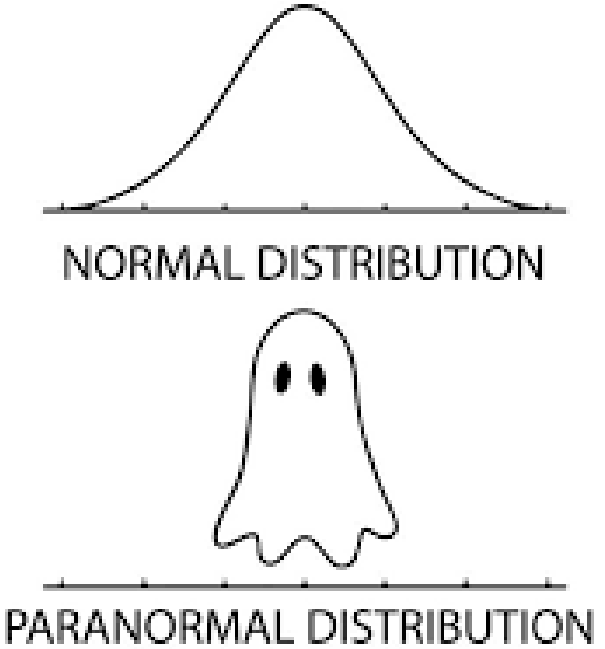
- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

TYPE OF NOT NORMAL DISTRIBUTION

1. Beta Distribution.
2. Exponential Distribution.
3. Gamma Distribution.
4. Log Normal Distribution.
5. Logistic Distribution.
6. Maxwell-Boltzmann Distribution.
7. Binomial Distribution
8. Poisson Distribution.
9. Skewed Distribution.
10. Symmetric Distribution.
11. Uniform Distribution.
12. Unimodal Distribution.
13. Weibull Distribution.

REASONS DATA FOLLOWS NOT NORMAL DISTRIBUTION

- Some data naturally follow non-normal distribution:
  - bacteria growth naturally follows an exponential distribution
  - number of accidents tends to fit a Poisson distribution
  - lifetimes of products usually fit a Weibull distribution.



NOT NORMAL DISTRIBUTION

Reasons data follow non-normal distribution:

a) Data collection: May be at fault.

b) Outliers: cause your data become skewed. Try removing any extreme high or low values and testing your data again.

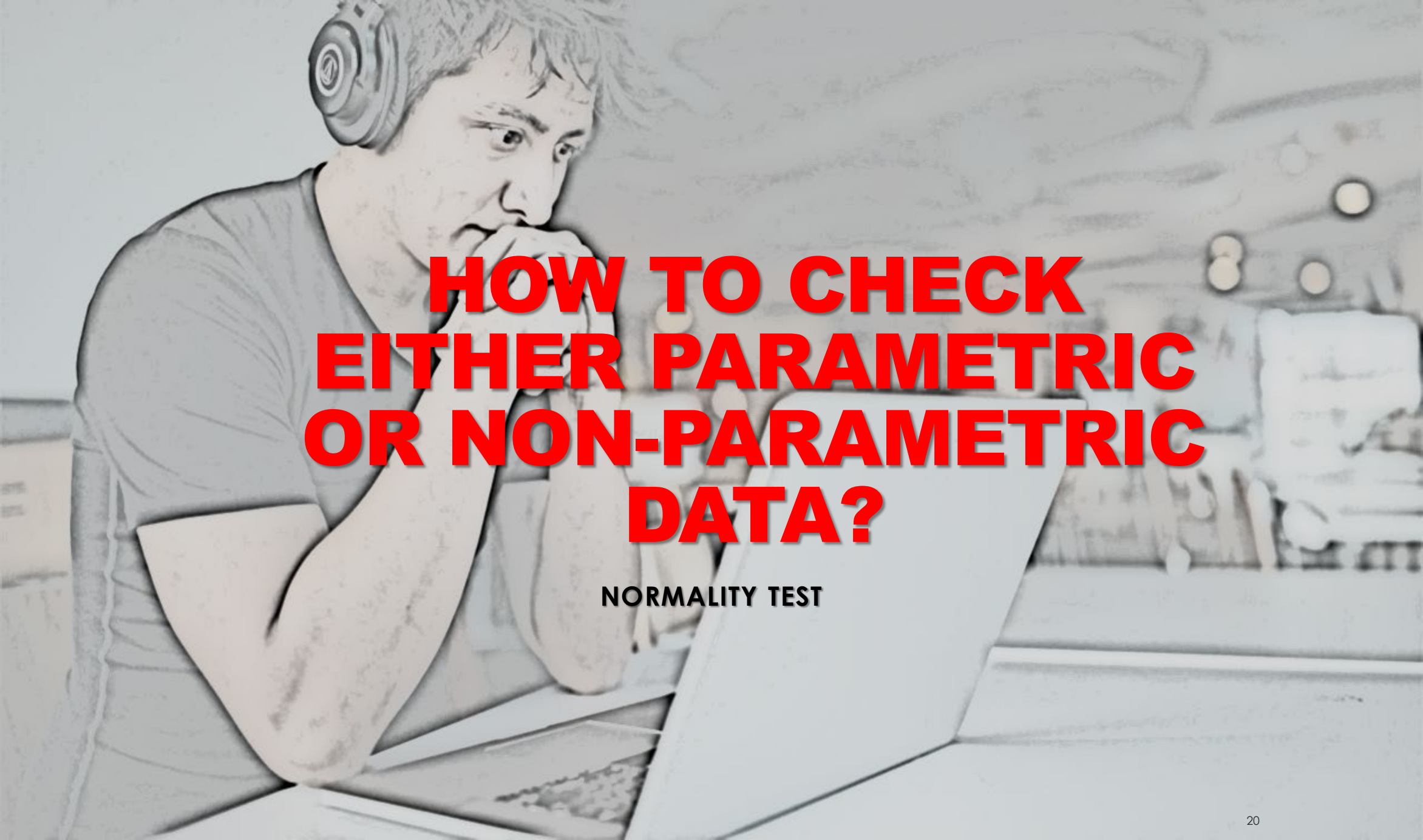
c) Multiple distributions may be combined in your data, giving the appearance of a bimodal or multimodal distribution.

d) Data may be inappropriately graphed.

- For example, if you were to graph people's weights on a scale of 0 to 1000 lbs, you would have a skewed cluster to the left of the graph.
- Make sure you're graphing your data on appropriately labeled axes.

e) Insufficient Data

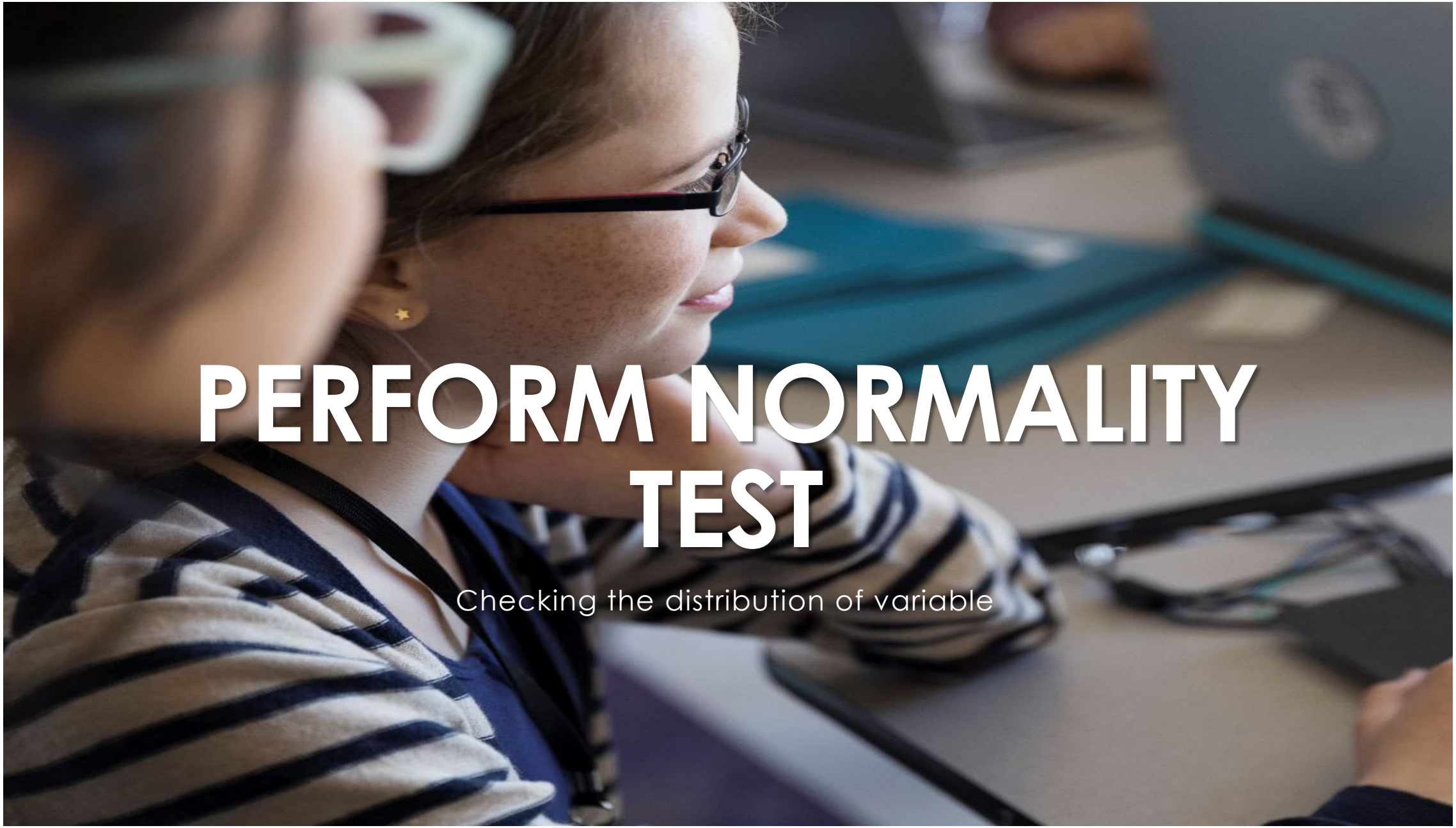
- For example, classroom test results are usually normally distributed. If you choose three random students and plot the results on a graph, you won't get a normal distribution.
- You might get a uniform distribution (i.e. 62 62 63) or you might get a skewed distribution (80 92 99).
- If you are in doubt about whether you have a sufficient sample size, collect more data.



HOW TO CHECK EITHER PARAMETRIC OR NON-PARAMETRIC DATA?

NORMALITY TEST

"BEFORE CONDUCTING STATISTICAL ANALYSIS ON uÜ ` ÇMÇ ÇM8 2 Ç , WE NEED TO TEST THE NORMALITY OF THE DATA. EITHER THE DATA IS A PARAMETRIC DATA (WITH NORMAL DISTRIBUTION) OR NON-PARAMETRIC DATA (WITH NOT NORMAL DISTRIBUTION)."



PERFORM NORMALITY TEST

Checking the distribution of variable

Hypothesis Statement for Normality Test:

H0 : Data is normally distributed

HA : Data is NOT normally distributed

1. Numerical methods: Conduct Statistical Test

STEP:

If LARGE SAMPLE ($n \geq 50$) use **KALMOGOROV SMIRNOV** test statistics.

If SMALL SAMPLE ($n < 50$) use **SHAPIRO WILK** test statistics.

CHECKING NORMALITY ASSUMPTION

Hypothesis test for Kolmogorov Smirnov for N > 50. If p-value is small than alpha, then H-null is rejected. If H-null is rejected then, data is NOT normally distributed.
Hypothesis test for Shapiro-Wilk Theory for N ≤ 50. If p-value is small than alpha, then H-null is rejected. If H-null is rejected then, data is NOT normally distributed.

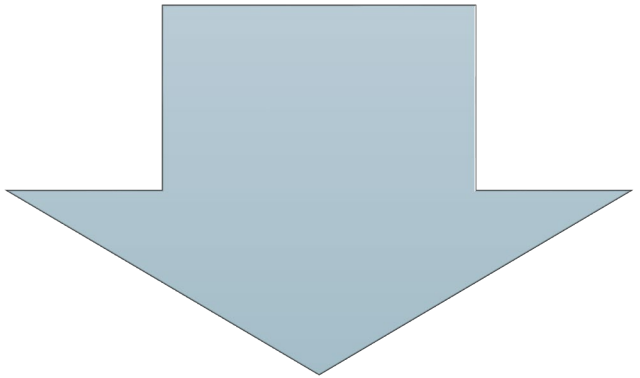
ONLY QUANTITATIVE DATA IS NEEDED FOR NORMALITY CHECKING

2. Visualization methods: Normal QQ Plot & Histogram

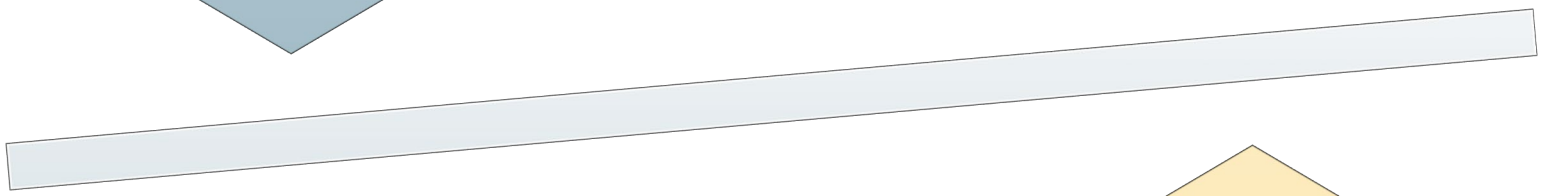
Normal Q-Q Plots :

If the data are normally distributed, the data points will be close to the diagonal line. If the data points stray from the line in an obvious non-linear fashion, the data are not normally distributed.

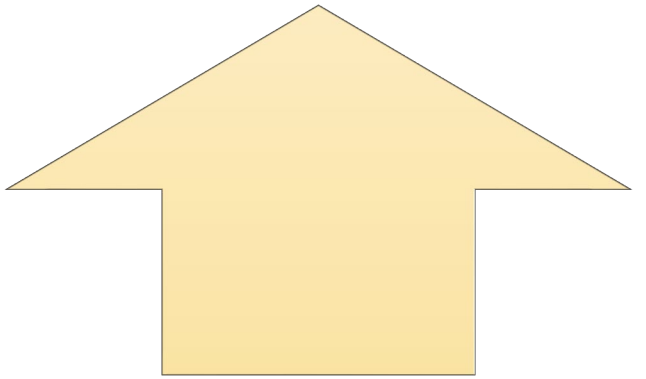
CHECKING DISTRIBUTION OF VARIABLE: NORMALITY TEST



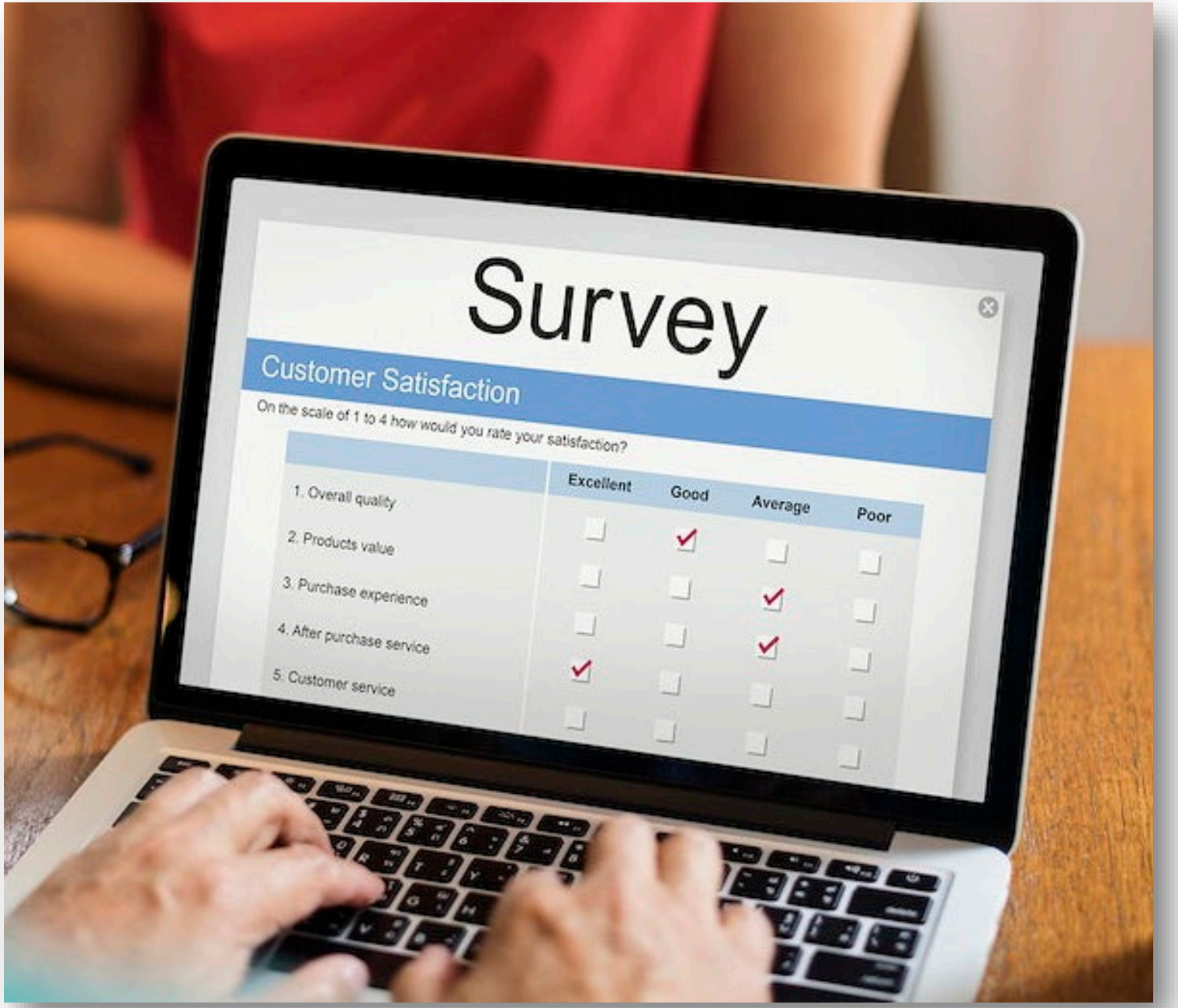
If distribution of a variable, follow NORMAL DISTRIBUTION USE PARAMETRIC TEST for statistical test



If distribution of a variable, follow NOT-NORMAL DISTRIBUTION USE NON-PARAMETRIC TEST for statistical test



PERFORM NORMALITY TEST WITH SPSS



LET'S PRACTICE

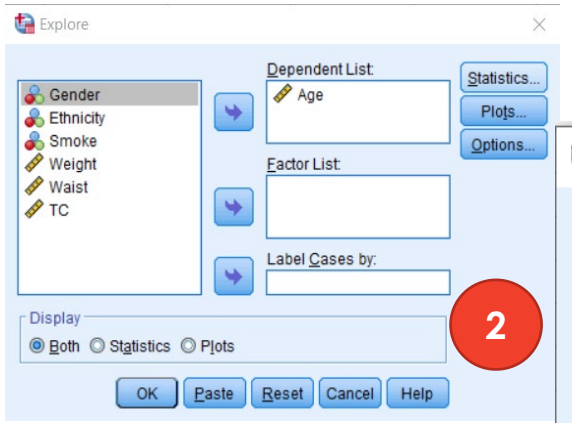
A survey on Health for 350 respondents at Ampang Community is conducted. The survey include the information on respondent's gender, age, ethnicity, smoking or not smoking, weight, waist measurement and cholesterol level. These data are in the file **HEALTH SURVEY.SAV**. Perform normality test to test the distribution of respondent **age**.

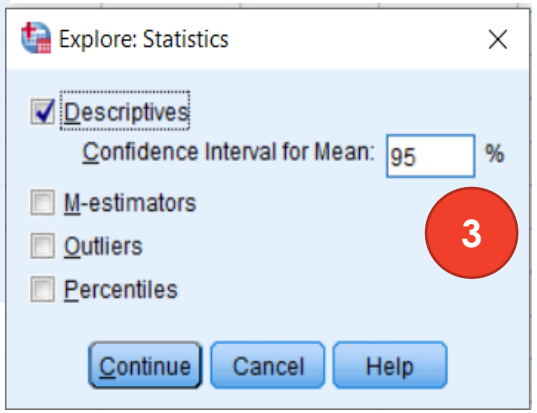
Test Normality using Numerical Method

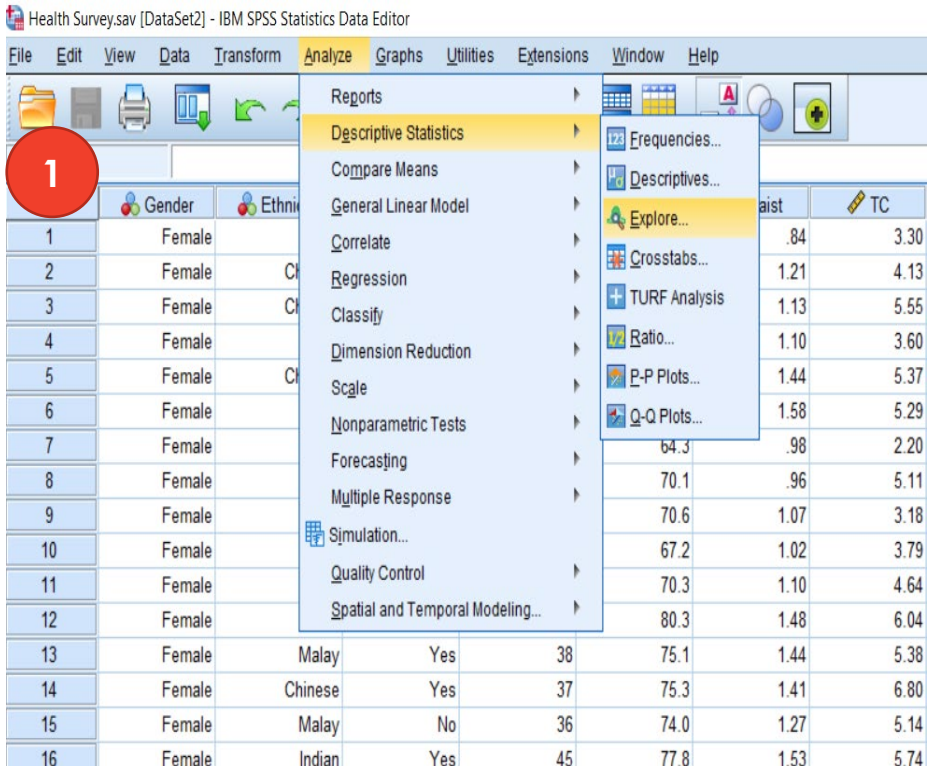
Data Set: Health Survey.sav

1. Analyze > Descriptive Statistics > Explore..

2. Select Age as Dependent, click Statistics

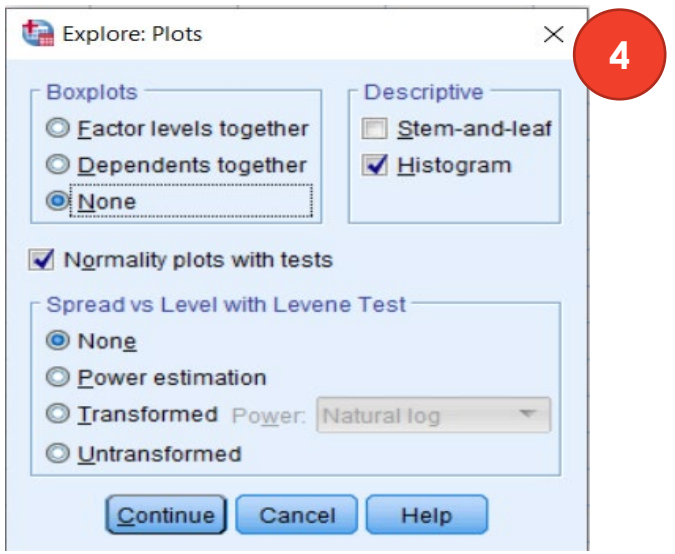






3. Statistics > Tick Descriptive .

4. Plots > Tick Histogram and Normality Plots with test



Test Normality using Numerical Method: OUTPUT

Data Set: Health Survey.sav

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age	350	100.0%	0	0.0%	350	100.0%

1

Descriptives

			Statistic	Std. Error
Age	Mean		40.00	.359
	95% Confidence Interval for Mean	Lower Bound	39.29	
		Upper Bound	40.71	
	5% Trimmed Mean		39.99	
	Median		39.75	
	Variance		45.000	
	Std. Deviation		6.708	
	Minimum		18	
	Maximum		58	
	Range		40	
	Interquartile Range		9	
	Skewness		-.002	.130
	Kurtosis		.103	.260

To summarize:
Mean = 40.00 years which is near to value of Median = 39.75 years
Skewness = -0.002 (between -0.5 and 0.5) for fairly symmetrical
Kurtosis = 0.10 which is near to 0 for normal distributed

2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.027	350	.200*	.996	350	.434

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Hypothesis Statement for Normality Test:

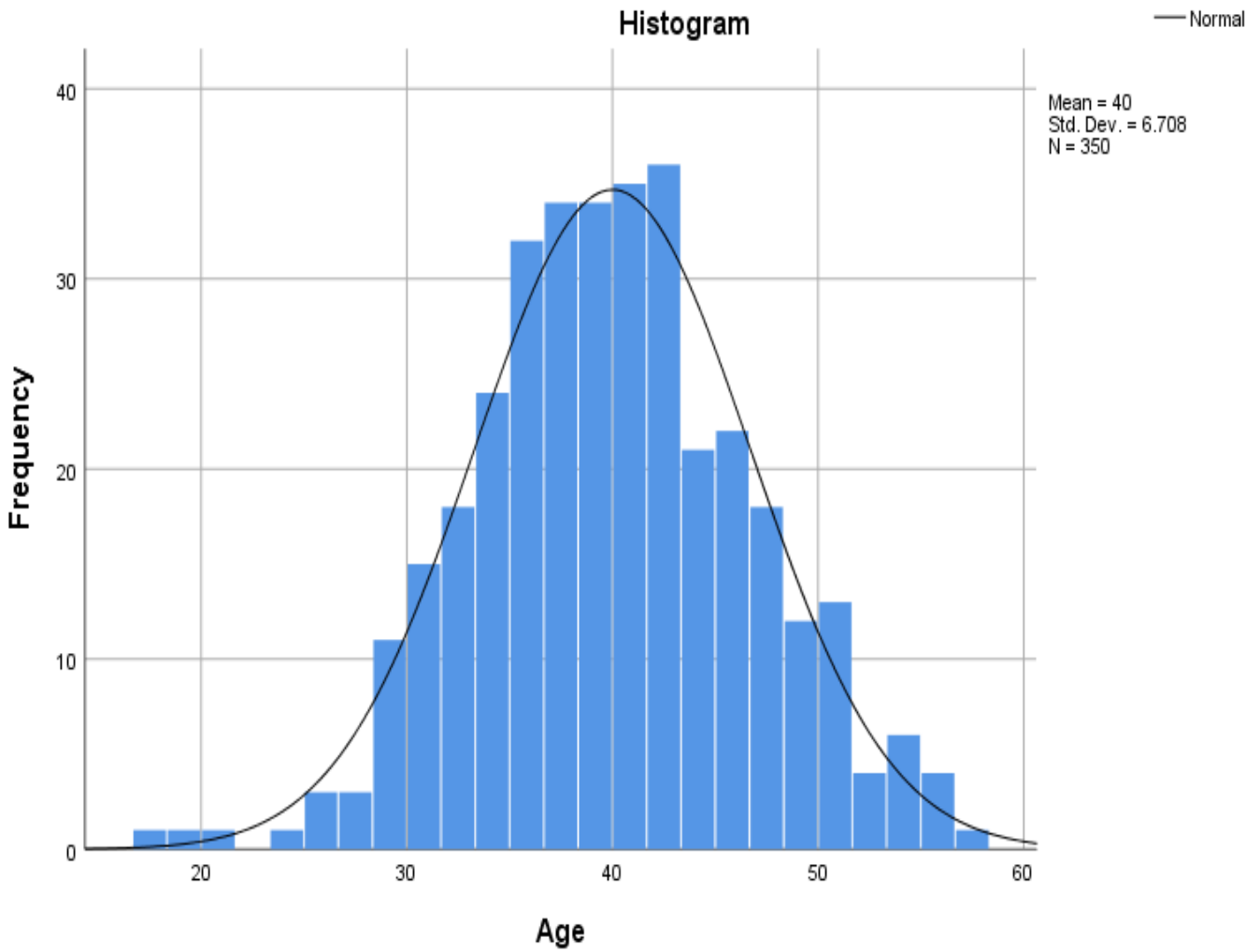
H_0 : Data is normally distributed
H_A : Data is NOT normally distributed

To summarize:
N = 350 samples therefore take Kolmogorov result:
P-value is 0.2 > ($\alpha = 0.05$), thus H-null is ACCEPTED. It indicate that data is normally distributed.

Test Normality using Visual Method: OUTPUT

Data Set: Health Survey.sav

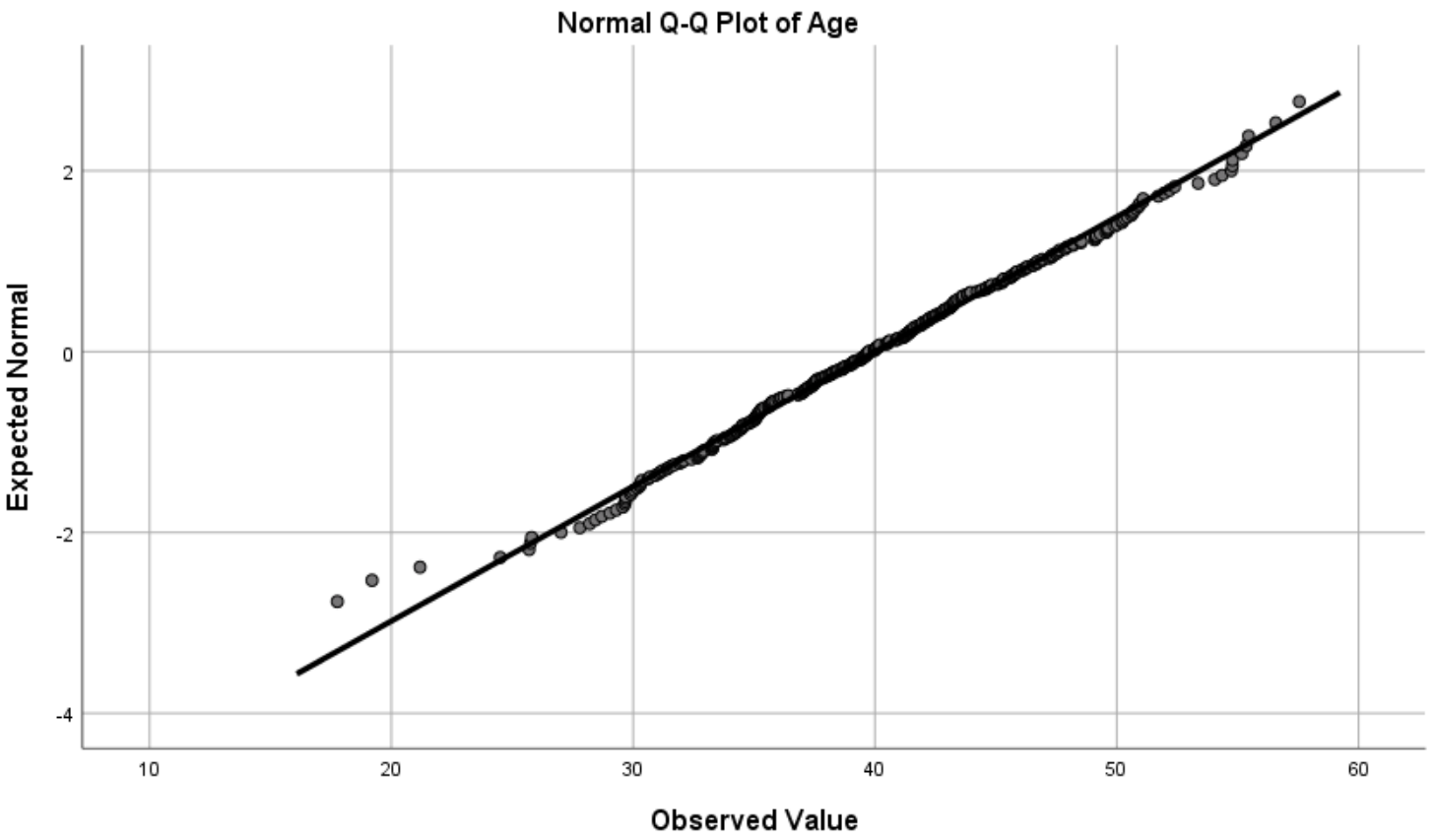
1



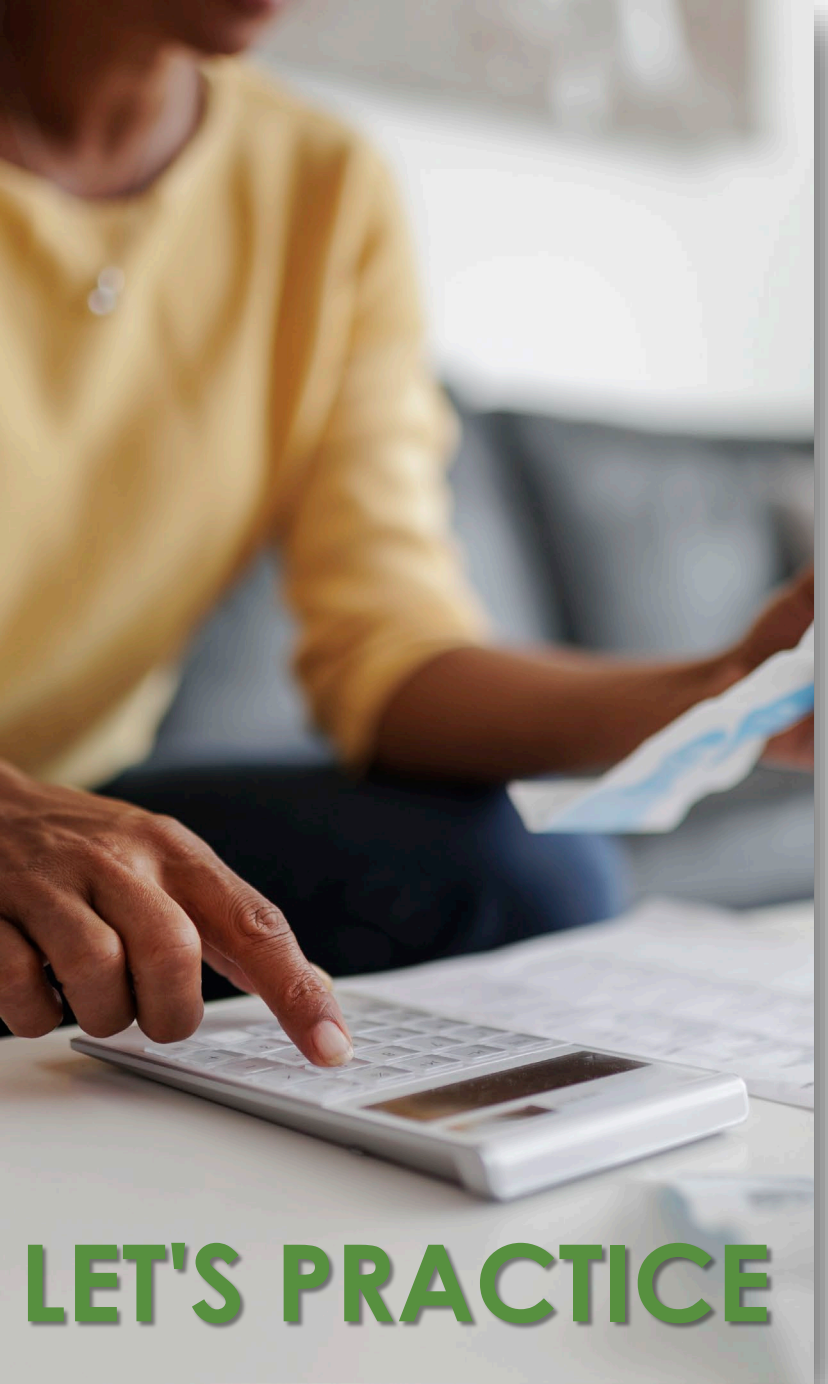
Test Normality using Visual Method: OUTPUT

Data Set: Health Survey.sav

2



data are normally distributed; the data points will be close to the diagonal line.



LET'S PRACTICE

Problem Statement:

A hospital's financial department is analyzing past-due medical bills to understand the distribution of patient payment delays. A financial officer claims that **past-due medical bill amounts are normally distributed**, which would allow them to use parametric statistical methods for financial planning and risk assessment.

However, other analysts suggest that **medical bill distributions are often skewed**, meaning a normality test is required before making further statistical assumptions.

- To test this claim, a random sample of **67 patients** was collected, recording the **past-due amounts** they owe.

Question:

Using the given dataset:

1. **Perform a normality test** on the past-due medical bill amounts.
2. **Interpret the results**—is the distribution normal or skewed?
3. **What implications** does this have for financial analysis in healthcare billing?
4. If the data is not normal, **what alternative statistical methods** could be used for further analysis?

Test Normality using Numerical Method: OUTPUT

Data Set: Healthcare_Past Due Bills.xlsx

1

Descriptives

			Statistic	Std. Error
Past_Due_Amount	Mean		27.23696	3.465329
	95% Confidence Interval for Mean	Lower Bound	20.31820	
		Upper Bound	34.15571	
	5% Trimmed Mean		24.52236	
	Median		16.96611	
	Variance		804.570	
	Std. Deviation		28.364940	
	Minimum		.624	
	Maximum		105.107	
	Range		104.483	
	Interquartile Range		30.806	
	Skewness		1.422	.293
	Kurtosis		1.111	.578

To summarize:
Mean = 27.236 and Median = 16.986
Skewness = 1.422
Kurtosis = 1.111

Hypothesis Statement for Normality Test:

H_0 : Data is normally distributed

H_A : Data is NOT normally distributed

2

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Past_Due_Amount	.200	67	<.001	.808	67	<.001

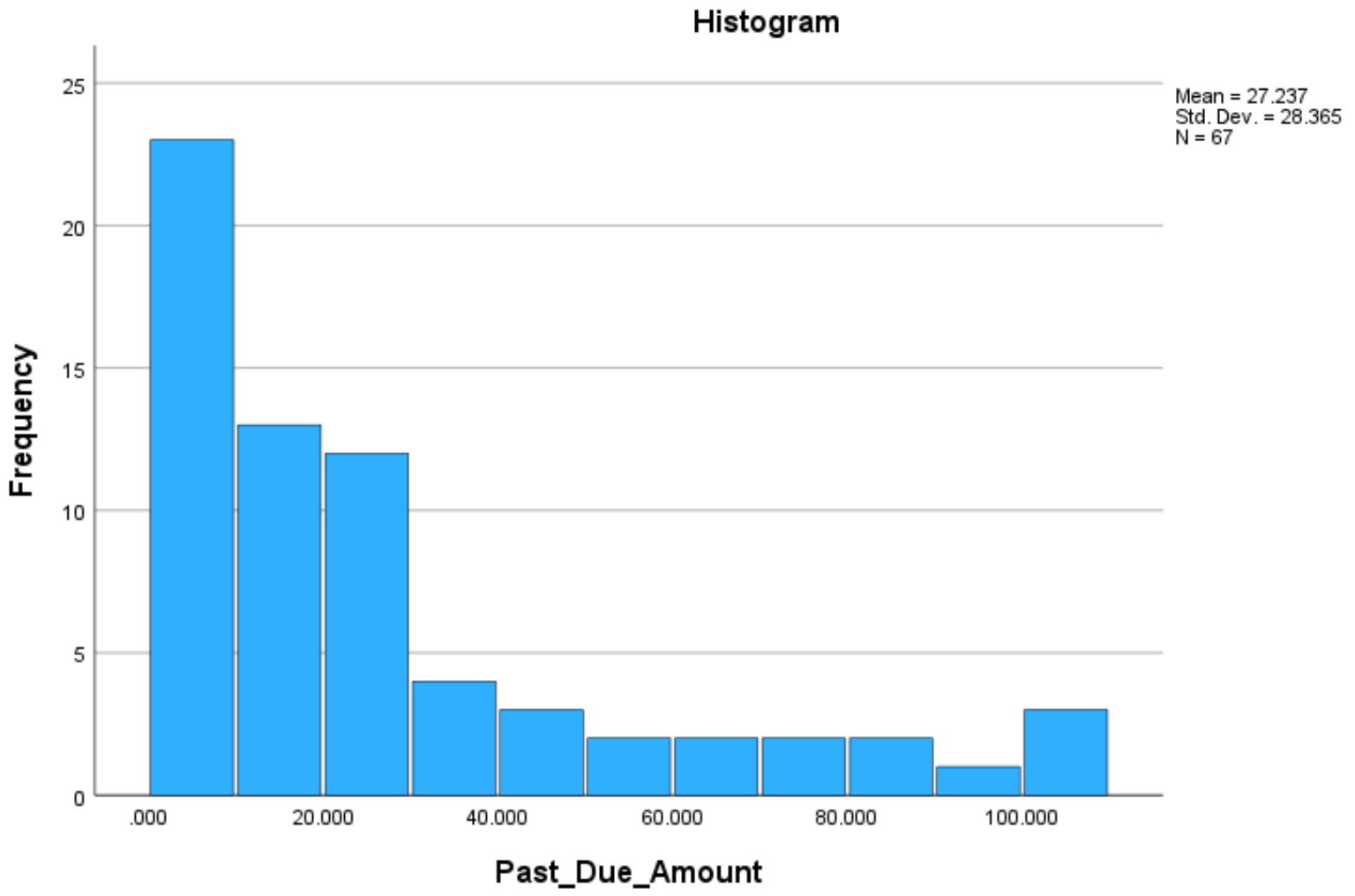
a. Lilliefors Significance Correction

To summarize:
N = 67 samples therefore take Kolmogorov result:
P-value is 0.000 < ($\alpha = 0.05$), thus H-null is REJECTED. It indicate that data is NOT normally distributed.

Test Normality using Numerical Method: OUTPUT

Data Set: Healthcare_Past Due Bills.xlsx

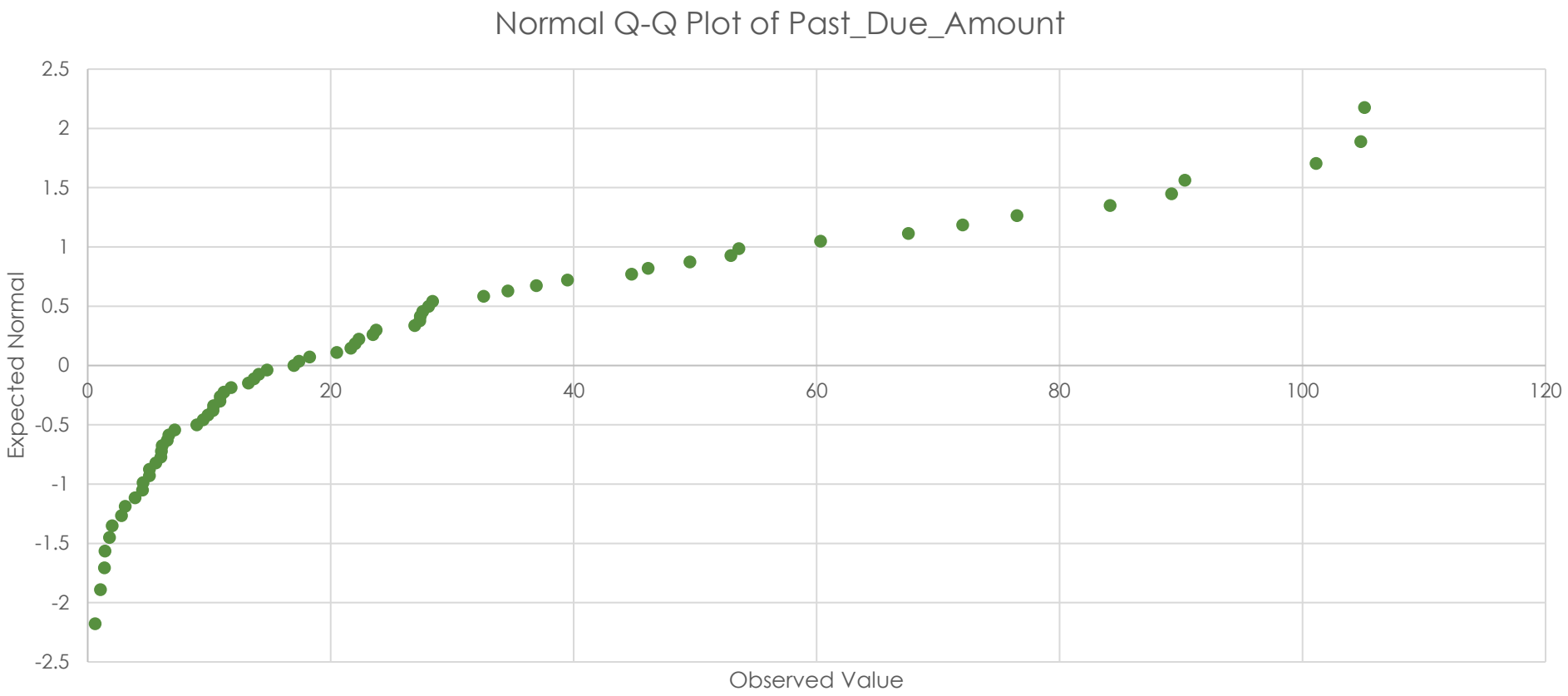
1



Test Normality using Numerical Method: OUTPUT

Data Set: Healthcare_Past Due Bills.xlsx

2



data are NOT normally distributed; majority of the data points is not close to the diagonal line.

INTERPRETATION FROM THE ANALYSIS

Implications for Healthcare Financial Analysis

- **Since the data is not normally distributed**, we **cannot use parametric tests** (e.g., t-tests, ANOVA) that assume normality.
- The hospital should use **non-parametric tests** (e.g., Mann-Whitney U test, Kruskal-Wallis test) for financial risk assessment.
- Predictive models should account for **skewness**, possibly using **log transformation** or robust statistical methods.

Alternative Statistical Methods for Non-Normal Data

- **Log Transformation**: Applying a logarithmic scale can reduce skewness.
- **Non-Parametric Tests**: Instead of standard t-tests, use **Mann-Whitney U** for comparisons.
- **Median-Based Metrics**: Since the mean is affected by skewness, using the **median** as a central measure is preferable.
- **Bootstrapping**: Resampling techniques can provide more reliable confidence intervals.

Conclusion:

- The **hospital's assumption of normality is incorrect**. Future financial planning should account for **skewed data** and use appropriate **non-parametric** or **transformative** statistical methods.

STATISTICAL METHOD FOR PARAMETRIC AND NON-PARAMETRIC?

PARAMETRIC

- Data follows a **normal distribution**.
- Descriptive Analysis:
  - Measure of Central Tendency → Mean
  - Dispersion → Standard Deviation
  - Shape → Bell Shape
- Statistical test:
  - T-test (one sample, Independent sample & Paired Sample)
  - F-test
  - Z-test
  - ANOVA
  - Pearson Correlation

NON-PARAMETRIC

- Data follows **not normal distribution**.
- Descriptive Statistics:
  - Measure of Central Tendency → Median
  - Dispersion → IQR
  - Shape → Non Bell Shape
- Statistical test:
  - Kolmogorov-Smirnov Test (KS test)
  - Mann-Whitney U test
  - Wilcoxon Signed Rank
  - Kruskal Wallis
  - Median test
  - Spearman Correlation

TEST YOUR UNDERSTANDING

Problem Statement:

A pharmaceutical company is analyzing **prescription fill times** (**in days**) to determine how long patients take to fill their prescriptions after receiving them. The finance and operations teams assume that the **fill times follow a normal distribution**, which would allow them to use parametric statistical methods for inventory forecasting and demand planning.

However, analysts suggest that **prescription fill times may be skewed**, as some patients delay refilling their prescriptions significantly. Before making financial decisions based on this data, a **normality test** must be performed.

- A random sample of **200 patients** was collected, recording the number of **days taken to fill their prescriptions**. **Data set** (Prescription_Fill_Time_Data.xlsx)
- Question:
  1. Perform a normality test on the prescription fill times
  2. Interpret the results—Is the distribution normal or skewed?
  3. What are the implications of a non-normal distribution for pharmaceutical supply chain planning?


THANK YOU