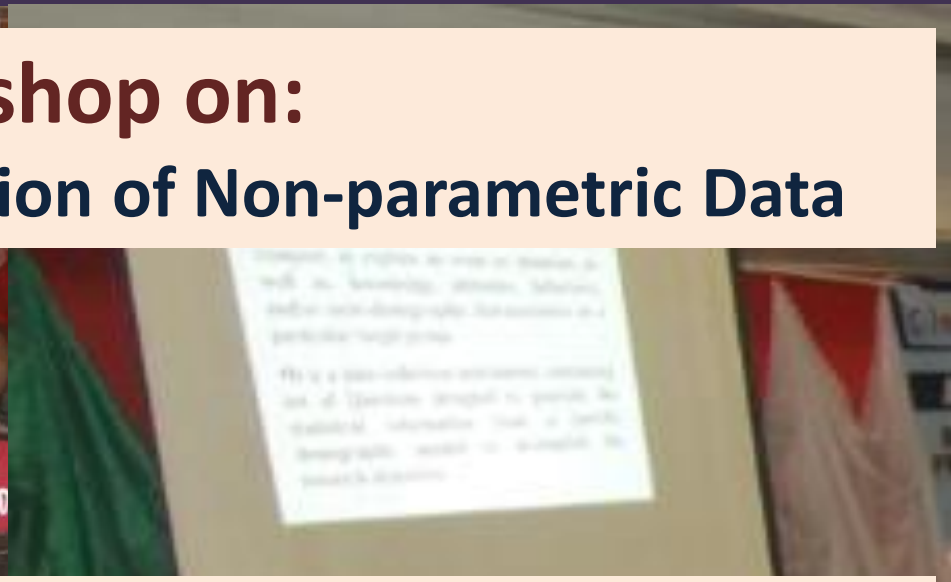
DOI: 10.33515/cint.sld/2023/001

C-International Tutors

Committed to bringing knowledge and skills to your doorsteps



Workshop on:

Analysis and Interpretation of Non-parametric Data

Module 1:

Introduction to Non-parametric Tests and Test of Normality

https://cintarch.com/tutorials-videos/

https://www.facebook.com/cintarch.tutors/

cintarch.tutors@gmail.com

https://www.instagram.com/cintarchtutors/?hl=en

https://twitter.com/CInternational6

https://www.youtube.com/channel/UCPEhxtU4B3Tu0PtPyMGBU8Q?view_as=subscriber

Introduction to Non-parametric Tests and Test of Normality

By

Prof. Awosan K.J.

Learning outcomes

After this session, you will know:

- the common types of non-parametric tests
- the properties of normal and skewed distributions
- the common tests of normality
- how to transform a non-parametric to a parametric data

Introduction

- **Data** refer to the records of two or more observations, while the record of a single observation is called datum
- Any named group of records is called a **data set**
- An **observation** is an event that is seen to occur
- Data are also defined as facts about something that can be used in calculating, reasoning or planning
- Facts or figures, or information that is stored in hard or soft form, or used by a computer are also called data
- Although, statistically data is plural, grammatically data can be used as a singular or plural in writing and speaking

Introduction contd.

- **Data analysis** is defined as the process of systematically applying statistical and/or logical techniques to describe and illustrate, condense and recap, and evaluate data
- It has also been defined as the process of cleaning, transforming and modelling data to discover useful information for decision making
- The main purpose of data analysis is to extract useful information from data and use the information obtained for decision making

Non-parametric tests

- **Non parametric tests** are methods of statistical analysis that do not require a distribution to meet the assumptions for a parametric test to be analyzed; as a result of this, they are sometimes called distribution free tests

- The indications for using non-parametric tests include:

➢ When the underlying data do not meet the assumptions about the population sample; for example, **when the data set is not normally distributed**

Non-parametric tests contd.

- The indications for using non-parametric tests include contd.:

➢ When the sample size is small, in which case one may not be able to validate the distribution of the data, the only option therefore is to apply a non-parametric test

➢ When the data is nominal or ordinal (i.e., categorical data); whereas, parametric tests are suitable for analyzing only continuous data, non-parametric tests are suitable for analyzing nominal or ordinal data

Non-parametric tests contd.

- The types of non-parametric tests and their corresponding parametric tests are shown in the table below

Table 4.3 Parametric and non-parametric statistical tests

Nature of groups	Type of variables	Parametric test (Purpose of test)	Non-parametric test (Purpose of test)
One group	Quantitative	Pearson's correlation (Test for relationship)	Kendall's tau, or Spearman rho correlation (Test for relationship)
One group compared with a population	Quantitative	1 sample t-test, or Z test (Compare means)	1 Sample Wilcoxon signed-rank test, or Sign test (Compare medians)
Two independent groups	Quantitative	Independent (or Unpaired) t-test (Compare means)	Mann-Whitney U test, or Wilcoxon rank sum test (Compare medians)
Two related Groups	Quantitative	Paired t-test (Compare means)	Wilcoxon matched pair signed-rank test (Compare medians)
Three or more Independent groups	Quantitative	ANOVA (Compare means)	Kruskal-Wallis rank sum H test (Compare medians)
Three or more repeated measures in one group	Quantitative	Repeated measures ANOVA (Compare means)	Friedman test (Compare medians)

Source: Awosan (2020)

Normal (or Gaussian) distribution

- If in a population, only a few are underweight or overweight (i.e., in the two extremes), while the majority weigh close to the mean weight for the population, then the population's weight is said to be normally distributed

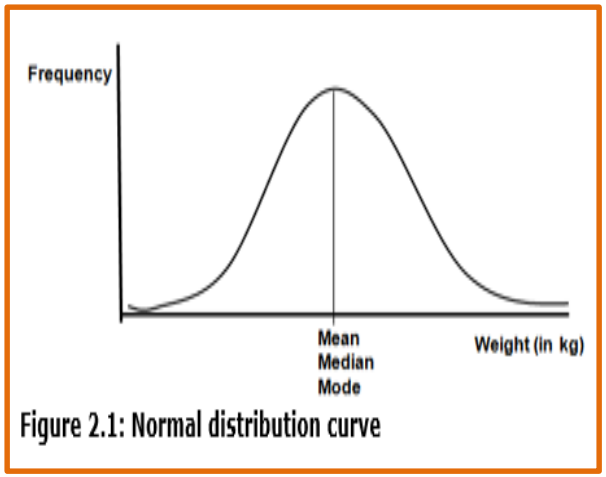


Figure 2.1: Normal distribution curve

Source: Awosan (2020)

- **In a normal distribution**:

  Mean = Median = Mode

Normal Distribution contd.

❑ Characteristics of a normal distribution curve

- The normal distribution curve is bell shaped
- It is ***symmetrical*** about the mean
- The curve on either side of the mean is a mirror image of the other
- The ***mean***, ***median*** and ***mode*** are all equal and located at the center of the distribution
- The curve is **unimodal** (i.e., has a single mode)
- The curve is continuous and have tails that are ***asymptotic*** (i.e., they approach but never touch the x-axis)

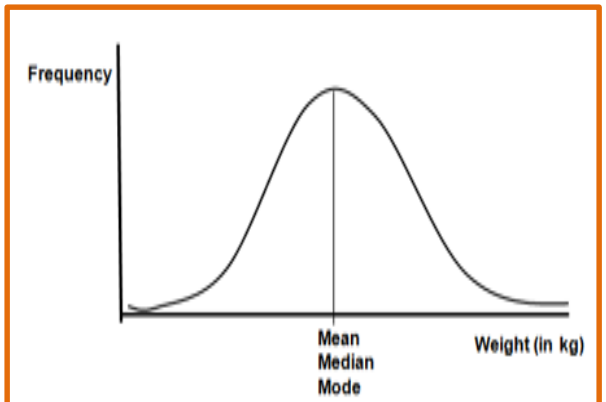


Figure 2.1: Normal distribution curve

Source: Awosan (2020)

Normal Distribution contd.

❑ Characteristics of a normal distribution curve contd.

- The total area under the normal distribution curve is 1

- Area corresponding to ±1SD will comprise 68.27% of the total area, ±2 SD will comprise 95.45% of the total area, and ±3 SD will comprise 99.73% of the total area. (68- 95- 99.7 rule)

- Also 95% of all values lie within 1.96 standard deviations, and 99% of all values lie within 2.58 standard deviations

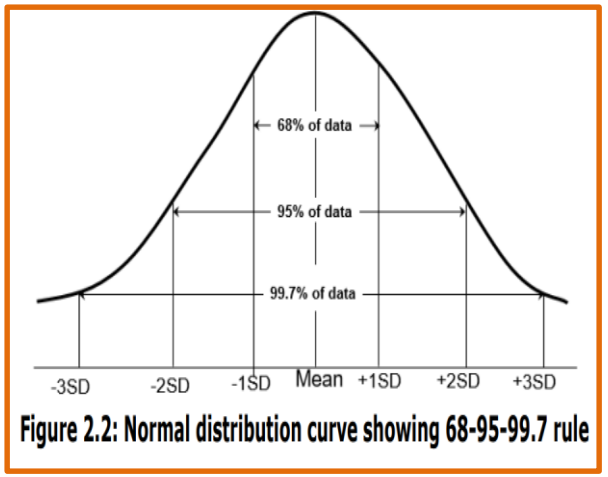


Figure 2.2: Normal distribution curve showing 68-95-99.7 rule

Source: Awosan (2020)

Left (or negatively) skewed distribution

- If the majority of the population are overweight, and only a few subjects in the population are underweight
- The distribution curve will show few people with underweight, and will have smaller area under the curve on the underweight (i.e., left or negative) side
- The distribution curve is said to be left or negatively skewed
- **In a Left skewed distribution**: **Mean < Median < Mode**

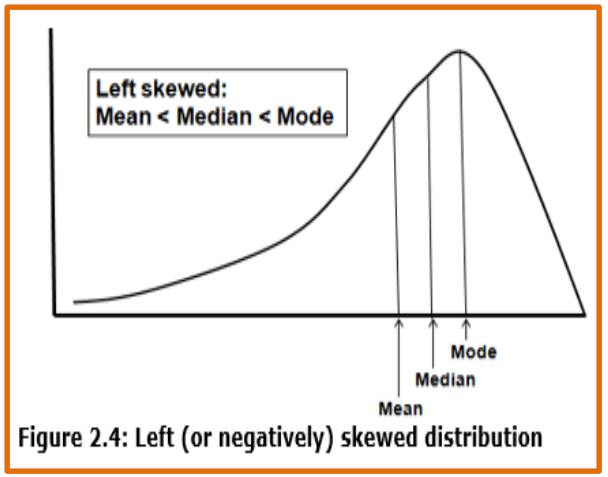


Figure 2.4: Left (or negatively) skewed distribution

Source: Awosan (2020)

Right (or positively) skewed distribution

- If the majority of the population are underweight, and only a few subjects in the population are overweight
- The distribution curve will show few people with overweight, and will have smaller area under the curve on the overweight (i.e., right or positive) side
- The distribution curve is said to be right or positively skewed
- **In a Right skewed distribution**: **Mean > Median > Mode**

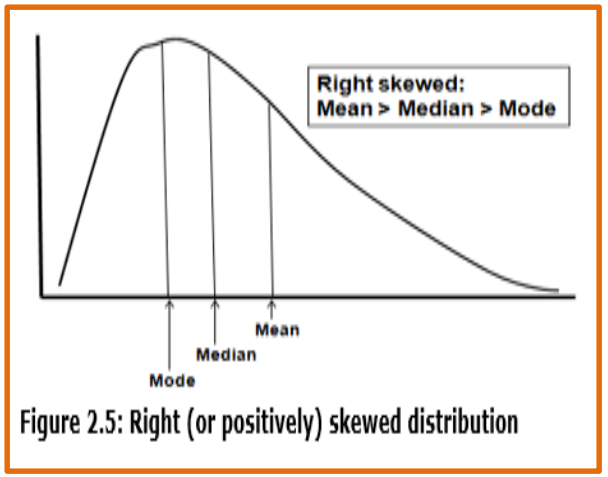


Figure 2.5: Right (or positively) skewed distribution

Source: Awosan (2020)

Test of normality

- Test of normality is performed to determine if a data set is normally distributed or not in order to know the appropriate statistical test to be used in analyzing it (i.e., parametric or non-parametric)
- This is because one of the assumptions for most parametric tests to be reliable is that the data set is approximately normally distributed
- However, data does not need to be perfectly normally distributed for the tests to be reliable
- The two main methods of testing for normality are graphical and numeric methods

Test of normality contd.

❑ Graphical method of testing for normality

- Testing for normality graphically can be done by creating the histogram, Q-Q plots, or box and whiskers plots of the dataset

❖ Histogram

- The curve of the histogram of a dataset that is normally distributed is symmetrical around its center (i.e., bell shaped with the right side of the center being a mirror image of the left side)
- It is unimodal (i.e., has only one mode or peak), and the center is located at its peak
- The curve is continuous and have tails that are asymptotic (i.e., they approach but never touch the x-axis)

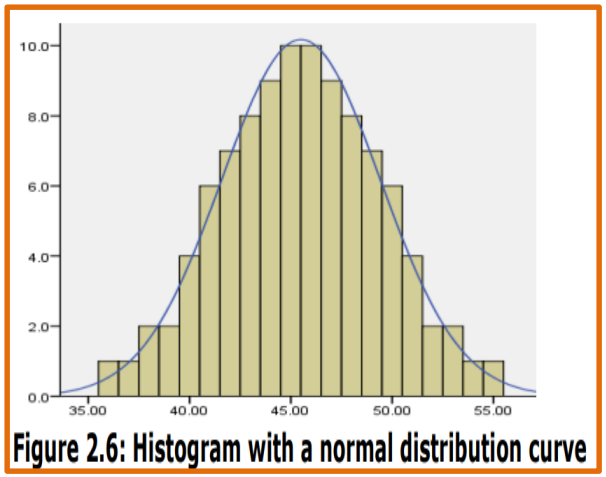


Figure 2.6: Histogram with a normal distribution curve

Test of normality contd.

- Graphical method of testing for normality contd.
  - Histogram contd.

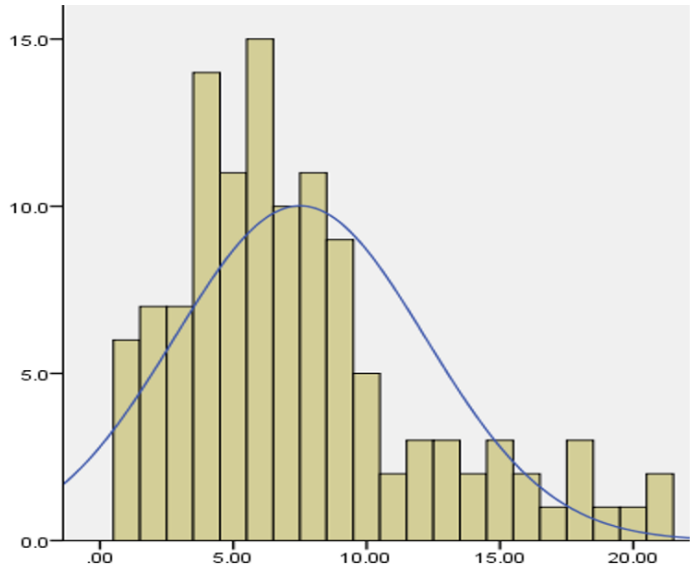


Figure 2.7: Histogram with a positively skewed distribution curve

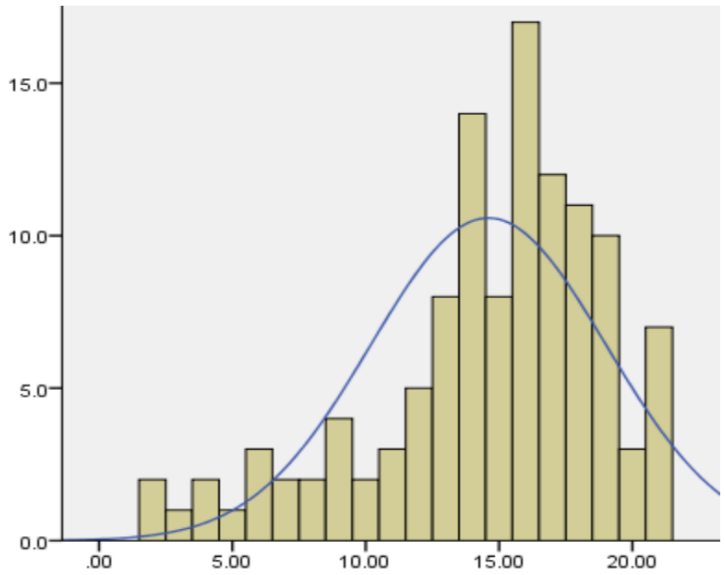


Figure 2.8: Histogram with a negatively skewed distribution curve

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Q-Q plots

- The Q-Q plot is a suitable graphical method of assessing normality for data sets with small sample sizes
- For a data set to be considered to be normally distributed the scatter plots should lie as close as possible to the line without any substantial deviation

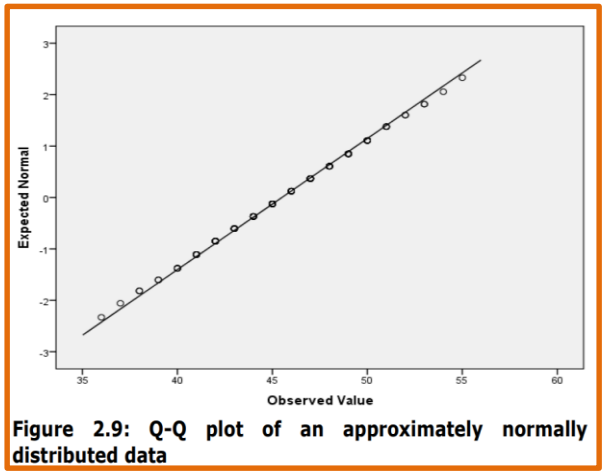


Figure 2.9: Q-Q plot of an approximately normally distributed data

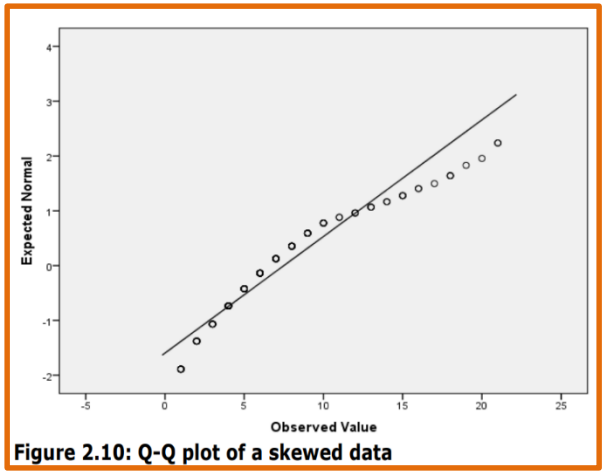


Figure 2.10: Q-Q plot of a skewed data

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Box and whiskers plot

- A box and whiskers plot is a type of graph used in summarizing a set of data measured on an interval scale
- It is useful in indicating whether a distribution is skewed, and whether there are potential unusual observations (i.e., outliers) in the data set
- Box and whiskers plots are also very useful when large numbers of observations are involved and when two or more data sets are being compared

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Box and whiskers plot contd.

- In a box and whiskers plot:

➢ The ends of the box are the upper and lower quartiles, so the box spans the inter-quartile range

➢ The median is marked by a horizontal line inside the box

➢ The whiskers are the two lines outside the box that extend to the highest and lowest observations (Figure 12.10)

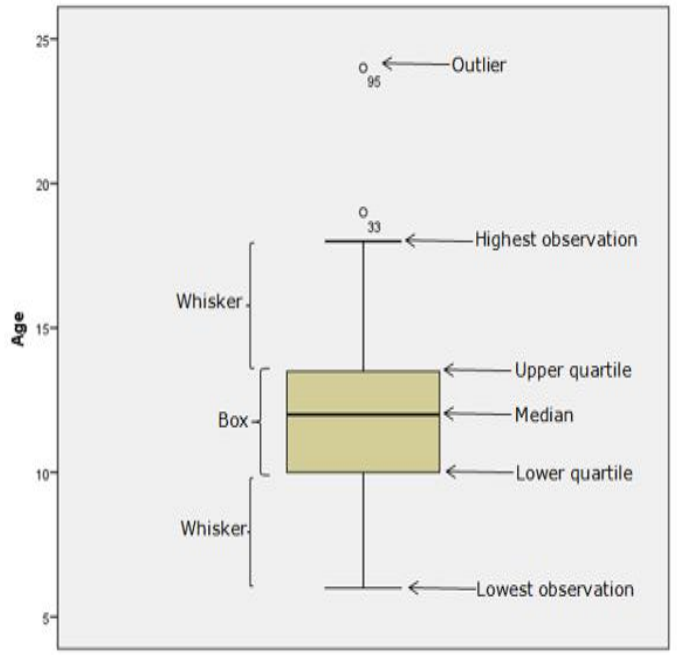


Figure 12.10: Box and whiskers plot

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Box and whiskers plot contd.

- The location of the median value in the box and whiskers plot is used to determine whether the distribution is normal, negatively skewed or positively skewed
- If the median is in the middle of the box, and the whiskers are roughly equal on each side, the distribution is "**Normal**" (Figure 12:10b)

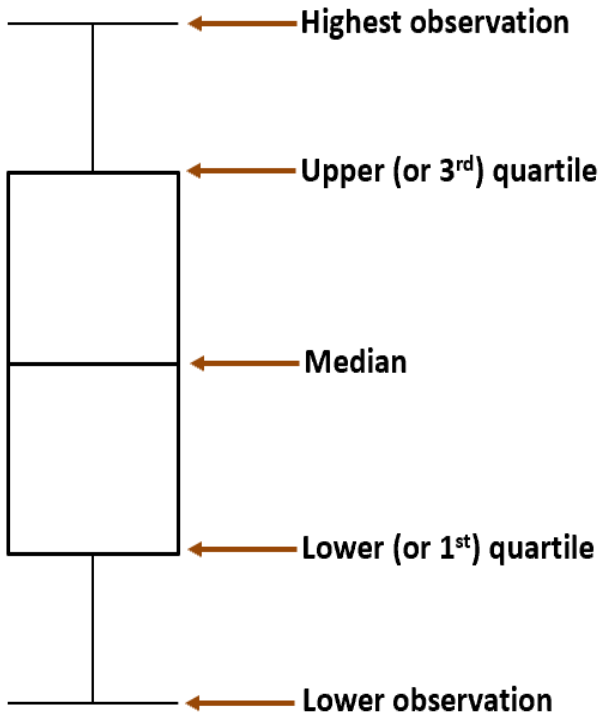


Figure 12.10b: Normal distribution

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Box and whiskers plot contd.

- If the median is closer to the bottom of the box and the whisker is shorter on the lower end of the box, the distribution is "**Right (or positively) skewed**" (Figure 12:10c)

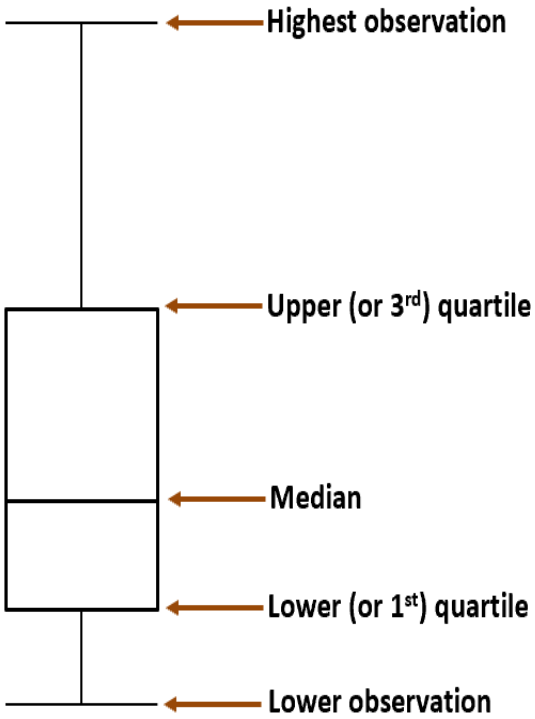


Figure 12.10c: Right (or positively) skewed

Test of normality contd.

❑ Graphical method of testing for normality contd.

❖ Box and whiskers plot contd.

- If the median is closer to the top of the box and the whisker is shorter on the upper end of the box, the distribution is "**Left (or negatively) skewed**"

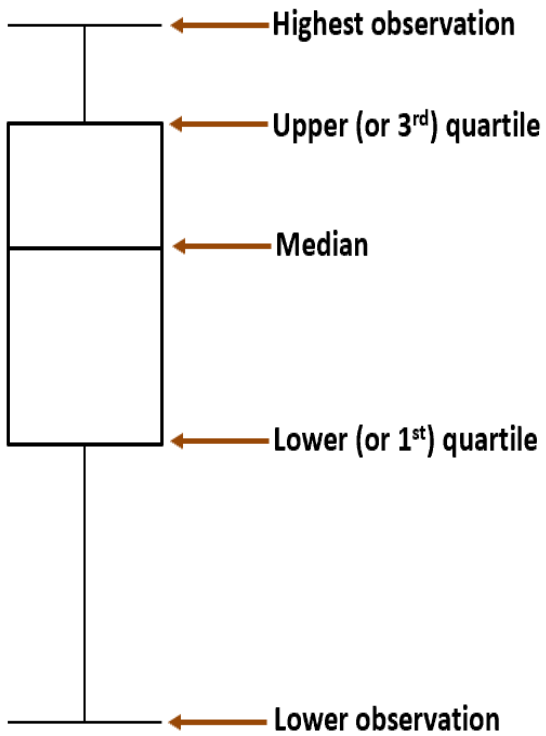


Figure 12.10d: Left (or negatively) skewed

Test of normality contd.

❑ Numeric method of testing for normality

- Testing for normality numerically can be done by **hypothesis testing** and by checking for the **skewness** and **kurtosis** of the data set

❖ Hypothesis testing

- Whereas, a data set may appear to be approximately normally distributed graphically, it may not really be normally distributed (as it may vary significantly from a normal distribution). Several tests are available for testing for whether or not a particular data set significantly differ from a normal distribution
- The most commonly used tests are the **Kolmogorov-Smirnov test** and the **Shapiro-Wilk's W test**, but the Shapiro-Wilk's test is preferred for small samples (less than 50).
- The **Null hypothesis** for the test of significance being performed is that "**the data is normally distributed**"

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Hypothesis testing contd.

- If the **p value is greater than 0.05** (i.e., given that α= 0.05), **Null hypothesis is not rejected**, which implies that **the data is normally distributed** (Figure 2.11). Both tests obtained p values > 0.05.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BP	.051	100	.200	.993	100	.876

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

a. Lilliefors Significance Correction

Figure 2.11: Test of normality of a normally distributed data

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Hypothesis testing contd.

- If the **p value is less than 0.05** (i.e., given that α= 0.05), **Null hypothesis is rejected**, which implies that **the data is not normally distributed** (Figure 2.12). Both tests obtained p values < 0.001 (i.e., < 0.05)

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Age	.144	118	.000	.911	118	.000

a. Lilliefors Significance Correction

Figure 2.12: Test of normality of a skewed data

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Skewness

- Skewness measures the asymmetry of a distribution
- A distribution is symmetric if it looks the same to the left and right of the center point
- In this case each half of the distribution curve is a mirror image of the other

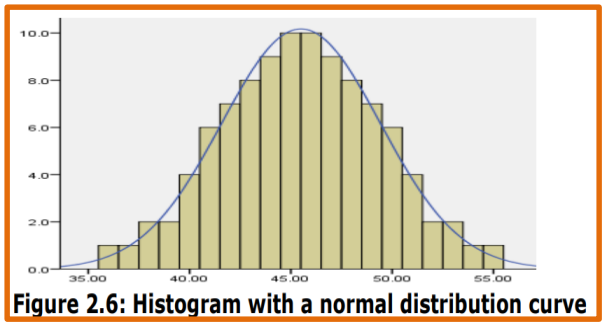


Figure 2.6: Histogram with a normal distribution curve

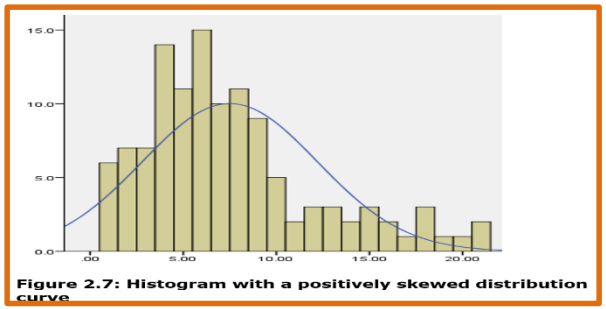


Figure 2.7: Histogram with a positively skewed distribution curve

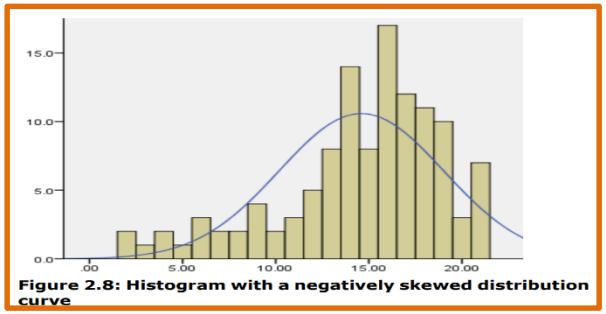


Figure 2.8: Histogram with a negatively skewed distribution curve

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Skewness contd.

- The skewness for a normal distribution is zero
- Any symmetric data (i.e., normally distributed data) should have a skewness close to zero
- A positive skewness value indicates that the data set is positively skewed or right skewed. This means that the right tail is long relative to the left tail
- A negative skewness value indicates that the data set is negatively skewed or left skewed. This means that the left tail is long relative to the right tail

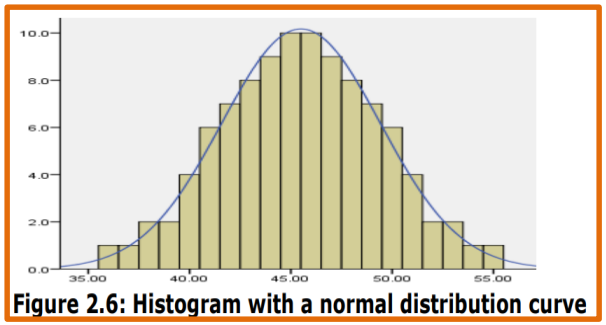


Figure 2.6: Histogram with a normal distribution curve

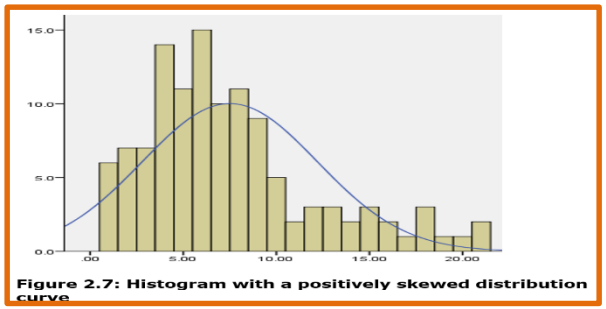


Figure 2.7: Histogram with a positively skewed distribution curve

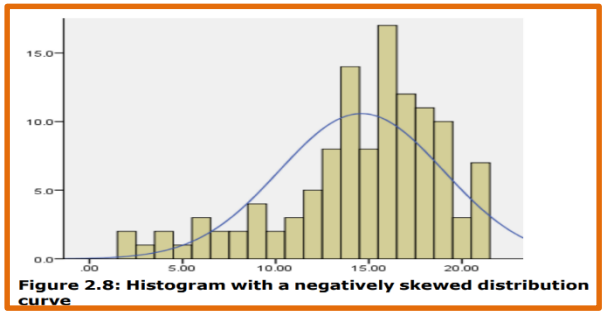


Figure 2.8: Histogram with a negatively skewed distribution curve

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Skewness contd.

- For a data with high mode, skewness is computed using the Pearson's first coefficient of skewness as:

$$Pearson's\ first\ coefficient = \frac{Mean\ - Mode}{Standard\ deviation}$$

- For a data with low mode or various modes, skewness is computed using the Pearson's second coefficient of skewness as:

$$Pearson's\ second\ coefficient = \frac{3\ (Mean\ - Median)}{Standard\ deviation}$$

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Skewness contd.

- Skewness is graded as follows:

Skewness	Classification
Lower than -1	Extremely negatively skewed
Between -1 and -0.5	Slightly negative skewed
Between -0.5 and +0.5	Nearly symmetric
Between +0.5 and +1	Slightly positively skewed
Greater than +1	Extremely positively skewed

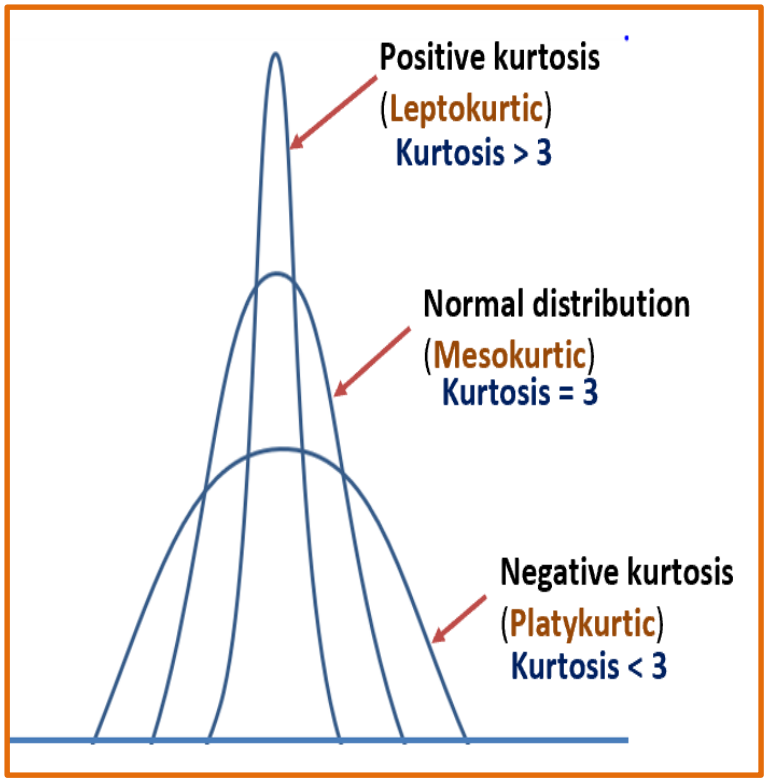
- Generally, a distribution is considered to be extremely skewed if the skewness is lower than -1 (negatively skewed) or greater than +1 (positively skewed); or if the skewness is double the standard error of the skewness or more

Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Kurtosis contd.

- Kurtosis measures the heaviness of a distribution's tail relative to a normal distribution
- It is used to determine whether a distribution is heavy-tailed or light-tailed relative to a normal distribution
- The kurtosis for a standard normal distribution is 3



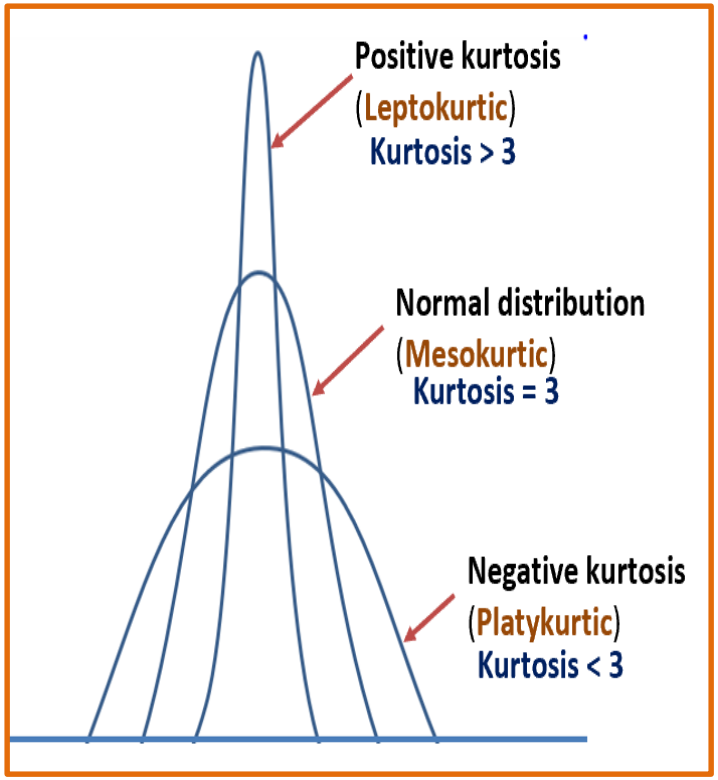
Test of normality contd.

❑ Numeric method of testing for normality contd.

❖ Kurtosis contd.

- The kurtosis for a standard normal distribution is 3, excess kurtosis is used to compare the kurtosis coefficient with that of a normal distribution
- Excess kurtosis can be **positive** (**Leptokurtic distribution**), **near zero** (**Mesokurtic distribution**), or **negative** (**Platykurtic distribution**)
- Excess kurtosis is computed as:

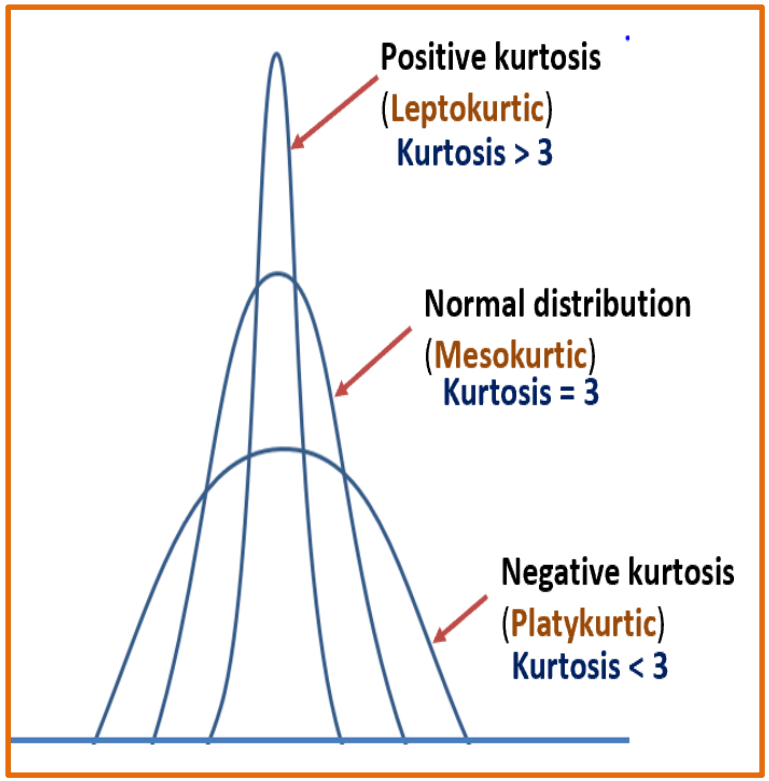
$$Excess\ kurtosis = Kurt\ - 3$$



Test of normality contd.

- Numeric method of testing for normality contd.
  - Kurtosis contd.
    - Excess kurtosis contd.

- **Leptokurtosis:** In this case there is a **positive excess kurtosis**. Leptokurtic distributions are fat-tailed, as they have many outliers
- **Mesokurtosis:** In this case, there is an **excess kurtosis of 0**. Normal distributions are mesokurtic
- **Platykurtosis:** In this case there is a **negative excess kurtosis**. Platykurtic distributions are thin-tailed, as they have few outliers



Analysis options for not normally distributed data

- The analysis options for a data set that is not normally distributed include the following:

➢ Use a non-parametric test: They are also called distribution free tests (this is the most valid option)

➢ Transform the dependents variable (such as taking the **log transformation** or the **square root transformation**) and repeat the normality check on the transformed data

➢ Use a parametric test under robust expectations; for example where the sample size is large enough, a parametric test can still be used instead of a non-parametric test (based on the central limit theorem)

Log transformation of data

- Log transformation is a data transformation method in which it replaces each variable x wit a log(x)
- It is used to transform a highly skewed variable into a more normalized dataset
- A logarithm can be defined with respect to a base (b) where the base b-logarithm of x is equal to y
- This is because x equals to the b to the power of y
- One can take any positive number as the base of the logarithm, but the most commonly used bases are 2 and 10

Log transformation of data contd.

- The relationship between y and $\log_b(x)$ is shown in the equations below:

$$y = log_b(x)$$

$$b^y = x$$

- **Base 2:**

➢ the base 2 logarithm of 4 is 2 (because $2^2 = 4$)

➢ the base 2 logarithm of 8 is 3 (because $2^3 = 8$)

- **Base 10:**

➢ the base 10 logarithm of 100 is 2 (because $10^2 = 100$)

➢ the base 10 logarithm of 1000 is 3 (because $10^3 = 1000$)

Log transformation of data contd.

- Although, log-transformation is widely used in biomedical and psychosocial research to deal with skewed data, there are serious problems in using it to deal with skewed data
- Despite the common belief that the log transformation can decrease the variability of data and make data conform more closely to the normal distribution, this is usually not the case
- Moreover, the results of standard statistical tests performed on log-transformed data are often not relevant for the original, non-transformed data

Log transformation of data contd.

- In some cases the transformed data can change from being Rt skewed to being Lt skewed or still Rt skewed , and in some cases the log transformation can even exacerbate the skewness of the data
- A more fundamental problem is that there is little value in comparing the variability of original versus log-transformed data because they are on totally different scales

Log transformation of data

- Another problem with log transformation is that It is also more difficult to perform hypothesis testing on log-transformed data

- It has been recommended that researchers should consider these limitations while using these traditional methods of dealing with skewed data. A safer option is to use newer analytic methods that are not dependent on the distribution of the data (i.e., non-parametric tests)

Test of Normality and Data Transformation in SPSS

- Test of normality and log transformation of skewed data in SPSS shall be covered in Module 2
- The datasets can be accessed through the links below:

Cintarch Dataset_Test of Normality 1_Normal distribution:
https://drive.google.com/file/d/1M8HiA1bsIl7IunWOekHBKE0ylomN-jBw/view?usp=sharing

Cintarch Dataset_Test of Normality 2_Positively skewed distribution:
https://drive.google.com/file/d/15LdvPLVzZFwpAvhmzHu3BNNQk_hv2fRs/view?usp=sharing

To access the videos, please visit: https://cintarch.com/tutorials-videos/

Test of Normality and Data Transformation in SPSS contd.

- The remaining dataset can be accessed through the link below:

Cintarch Dataset_Test of Normality 3_Negatively skewed distribution:
https://drive.google.com/file/d/1QyK6UyJSfG6Mxq5a86ztoNNPMNv7nTFP/view?usp=sharing

To access the videos, please visit: https://cintarch.com/tutorials-videos/

Further Reading

Awosan KJ (2020). Student Friendly Statistics for Health, Life and Social Sciences. Ikeja, Lagos: Somerest Ventures

Now On Sale

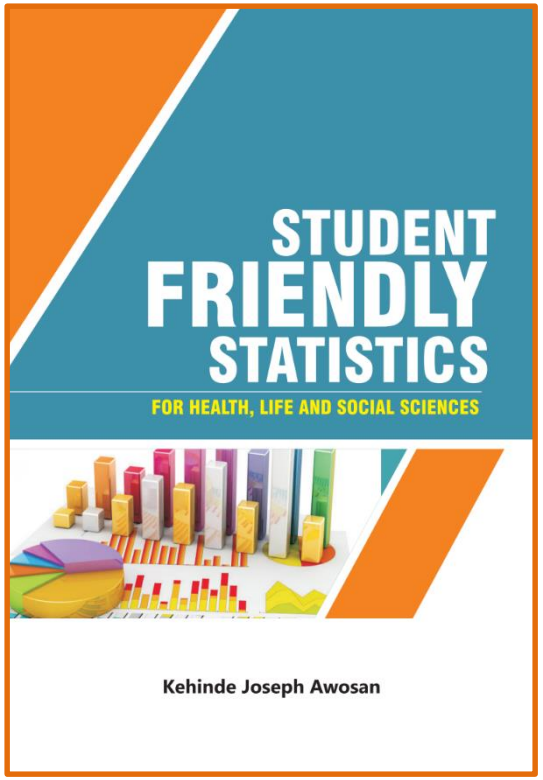
Student Friendly Statistics for Health, Life and Social Sciences

Please click on the link below to access the excerpts from the book:

https://cintarch.org/wp-content/uploads/erf_uploads/2022/02/CB_Excerpts-from-Student-Friendly-Statistics-for-Health-Life-and-Social-Sciences.pdf

To buy a copy or copies of the book online, please click on the link below to access the "ORDER FORM":

https://forms.gle/3J8dWn6Ng6oMKd2e8



C-International Archives' Journals

Please find the List of our Journals at:

https://cintarch.com/journals/

Please find our Recently Published Articles at:

https://cintarch.com/about-us/

Please Submit your Manuscripts at:

https://cintarch.com/submit-manuscripts/