THE BASIC APPLICATION OF BIOSTATISTICS TO BIOMEDICAL SCIENCE USING R PROGRAMMING

PRESENTER: OGBOLU MELVIN OMONE

OVERVIEW

- In biomedicine, the study of statistics is known as biostatistics.
- It involves the applications of control methods and pathophysiological modeling for data interpretation and presentation.
- This paper reveals the concept of data analysis as related to biostatistics and its methods which are useful for Statisticians with the use of R programming language.

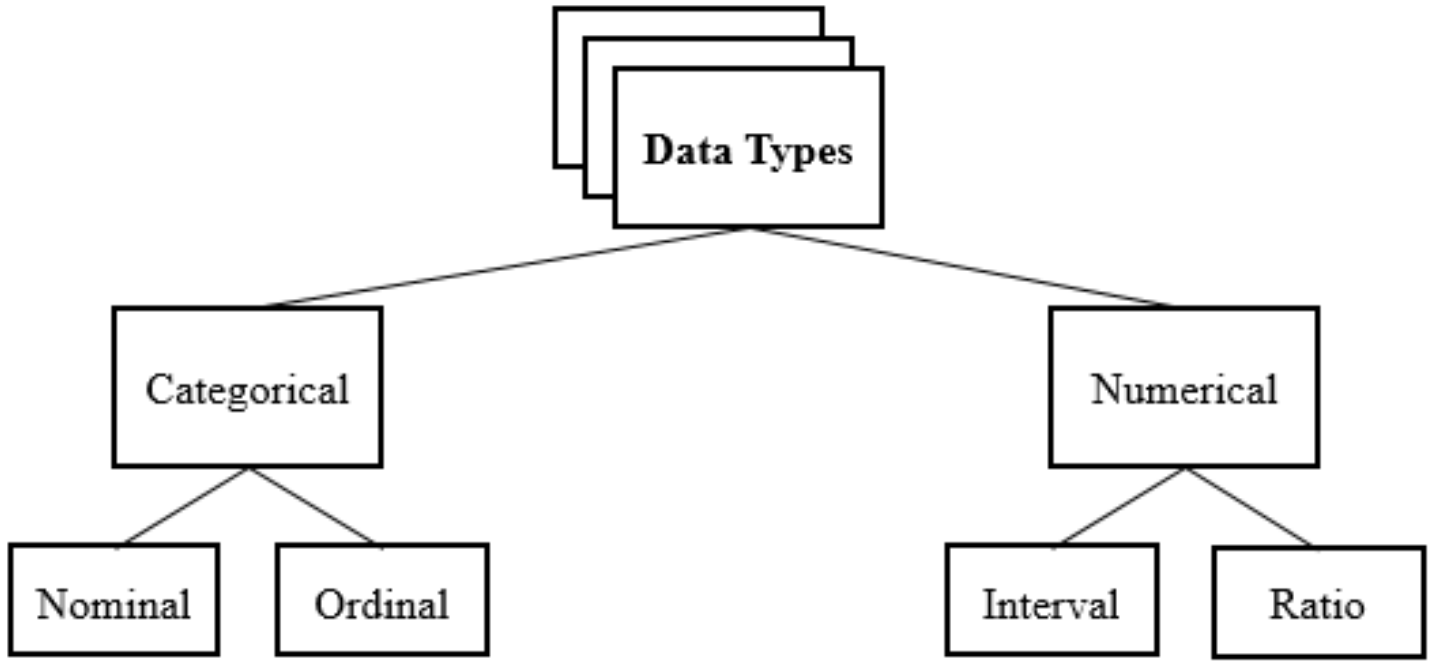
AREAS OF BIOSTATISTICS

- Applications of biostatistics, and
- The fundamental methods and techniques of biostatistics.

APPLICATIONS OF BIOSTATISTICS

- Statistics is the collection, organization, summarization, and analysis of raw data (mostly numerical data) in large quantities with the application of mathematical formulas, tools/methods, and software applications.
- Raw data are in form of data types. Which are;
  - Nominal data
  - Ordinal data
  - Interval data
  - Ratio data

Fig 1: Hierarchical structure of Data types



THE FUNDAMENTAL METHODS AND TECHNIQUES OF BIOSTATISTICS

- Statistics has two (2) major methods used for problem solving and for carrying out statistical tests. They are;
  - Descriptive method: This method explains basic features and the distribution of population measurements, whereas data types are investigated using certain statistical methods, such as estimates of central tendency (i.e. mean, mode, and median), correlation coefficient, and measures of variability (i.e. standard deviation and correlation coefficient)
  - Inferential method: It is used to express the level of certainty about estimates about a population which includes hypothesis testing, standard error of mean, regression analysis, and level of significance

TOOLS FOR DESCRIPTIVE STATISTICS

A. Measures or Estimates of Central Tendency

i. **Mean (M)**: It is commonly used for measuring central tendencies and for the calculation of averages. For exmaple: find the mean value of birthwt$bwt, we have;

> mean(birthwt$bwt) or with(birthwt, mean(bwt)) #we use this command to find the mean value of birthwt and bwt (mean of all numeric variables)

Output:
2944.587

ii. **Mode (Mo):** It reveals the most frequent/appearing value in a set of observation or a population. For example: find the group of children in the race group which appears the most, we have;

> table(birthwt$race) #we use this command to show the race column in a tabular form

Output:

Caucasian	Afro-American	Other
96	26	67

```
> max(table(birthwt$race)) #we use this command to find the mode by frequency
```

Output:

96

```
> names(sort(-table(birthwt$race)))[1] #we use this command to print the name of the mode
```

Output:

"Caucasian“

iii. **Median (Me):** It is an average value (the 50^{th} percentile) which separates the higher half of a sample from the lower. For example: find the median value of birthwt$bwt, we have;

```
> median( birthwt$bwt ) #we use this command to find the median value of birthwt and bwt
```

Output:

2977

B. *Measures of Variability*

i. **Standard Deviation (SD):** It depicts the variability of the examination of the **Mean (M)**. It is imperative to know that to find the value of SD, its square called variance must first be calculated. Hence, *Variance* = $\Sigma(x- \bar{x})\ 2/n$ *Or* $\Sigma(x- \bar{x})\ 2/n-1$. *Thus, SD* = $\sqrt{}$*variance*. For example: find variance before standard deviation;

```
> var(birthwt$bwt) #we use this command to find the variance of bwt
```

Output:
531753.5

```
> sd(birthwt$bwt) #we use this command to find the standard deviation of bwt
```

Output:
729.2143

ii. **Correlation Coefficient:** Correlation can be defined as the existing relationship between two variables. It is used to measure the degree (how strong) of linear relationship between two continuous variables. The mathematical formula for correlation coefficient is;

$$r = \frac{(\sum xy) - \frac{\sum x \sum y}{N}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right]\left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}} \ldots\ldots\ldots (1)$$

For example, the results for correlation is given below;

> cor(birthwt$lwt, birthwt$age) #Finding correlation between Low Weight and Age

Output:

0.1800732

> ggscatter(birthwt, x = "age", y = "lwt", add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Age of Children", ylab = "Children with Low Weight") #Graphical view

Output:

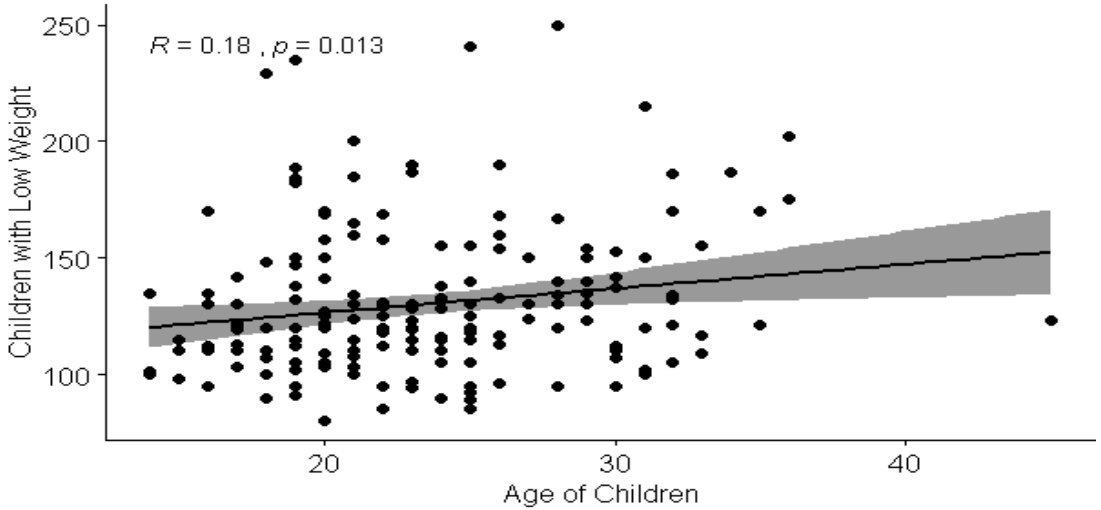


Fig 2: GGScatter for Correlation Coefficient between Age and Lwt using Pearson

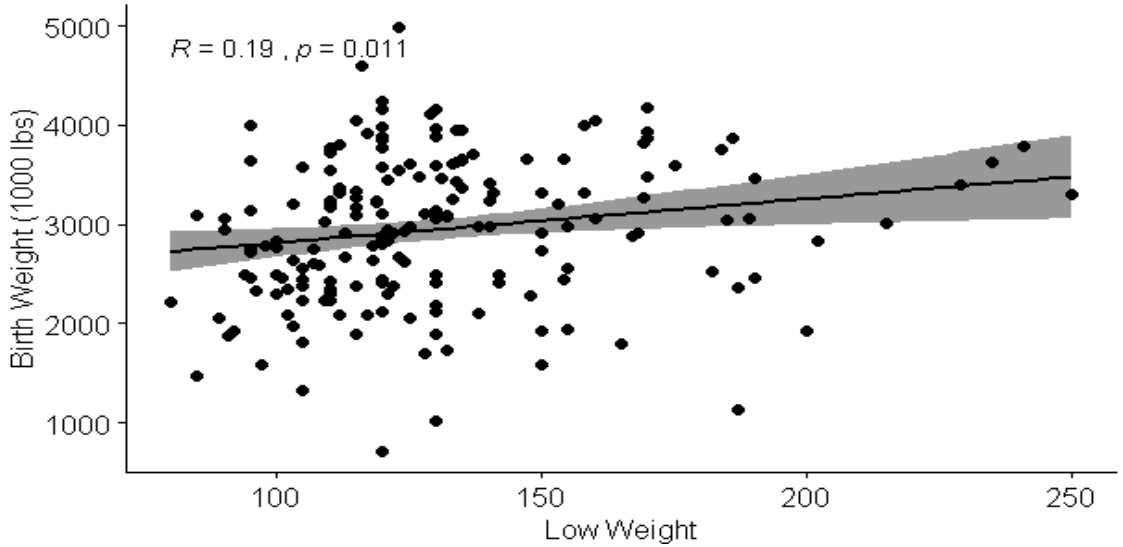


Fig 3: GGScatter for Correlation Coefficient between Bwt and Lwt using Pearson

TOOLS FOR INFERENTIAL STATISTICS

A. *Types of Distribution*

i. **Gaussian or Normal Distribution:** When data is symmetrically distributed on both sides of the Mean (M) and forms a bell-shaped (i.e. the standard normal distribution curve) in the frequency distribution plot, such distribution is called normal/Gaussian distribution.

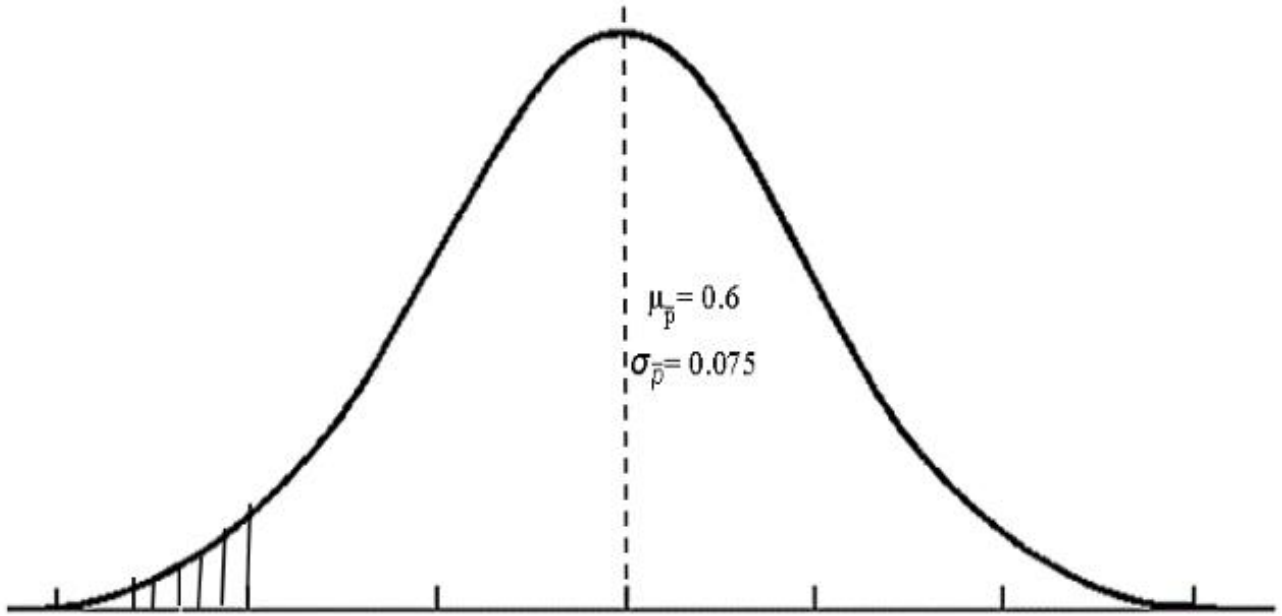


Fig 4: Normal Distribution Curve for One-sample

- In a normal distribution curve, the values of the **Mean (M), Mode (Mo),** and **Median (Me)** are usually the same within the population under investigation (i.e. the mean, mode, and median are all equal). For example, the results for normal distribution is given below;

```
> hist(birthwt$bwt) #Plots a histogram for birthwt$bwt
```

Output:

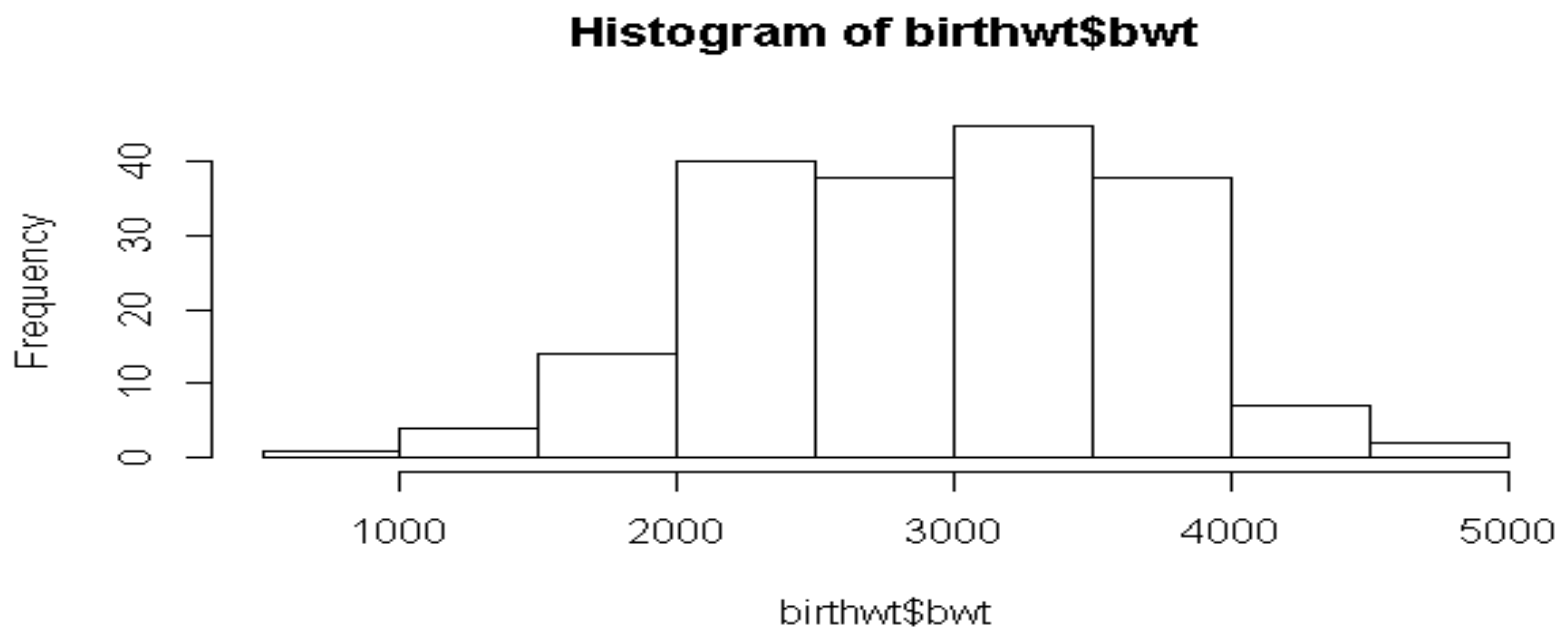


Fig 5: Normal Distribution for birthwt$bwt using Histogram

ii. **Non-Gaussian (non-normal) Distribution:** In this case, if the data is skewed on one side of the graph, then the distribution is called a non-normal distribution.

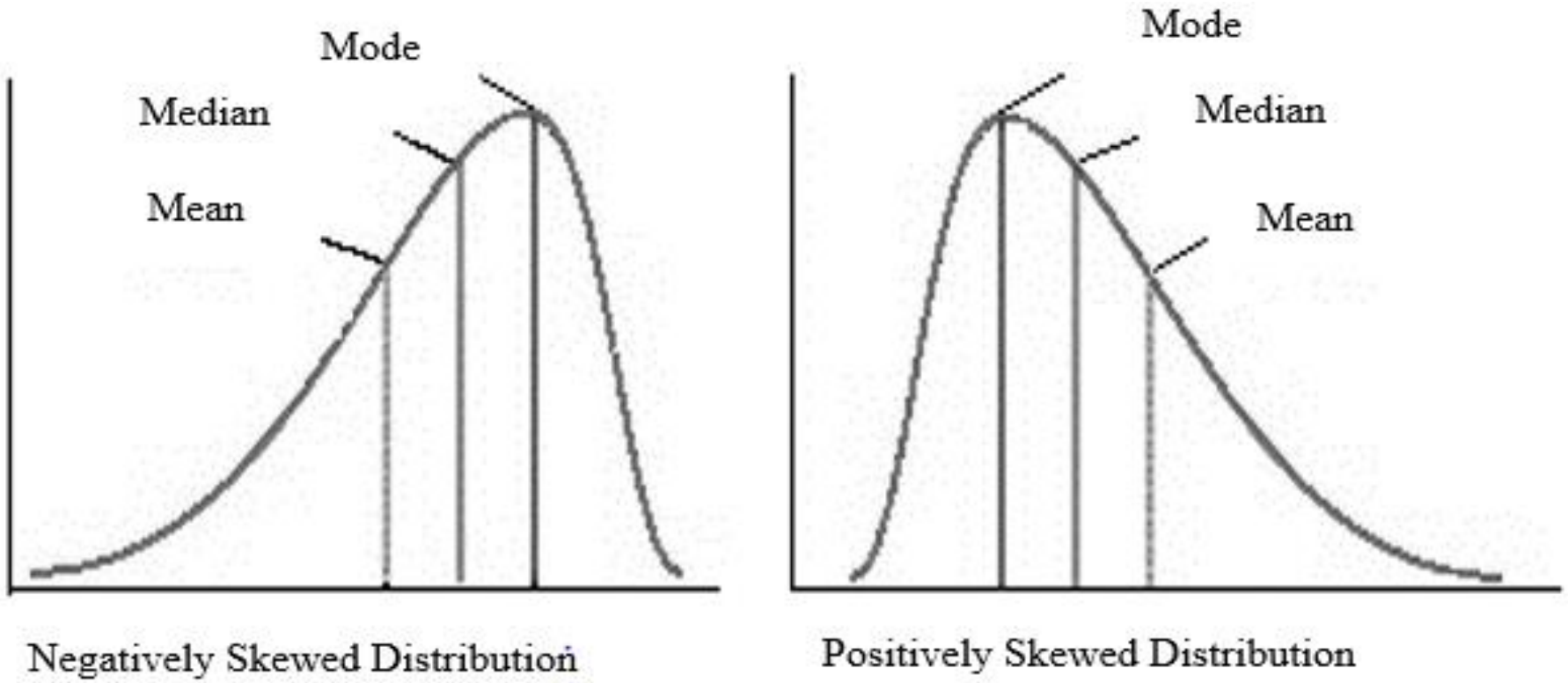


Fig 6: Negatively-skewed Distribution and Positively-skewed Distribution

- . For example: check the normality and distribution of birthwt$lwt and birthwt$bwt using a histogram;

  > hist(birthwt$lwt) #Plots a histogram for birthwt$lwt

  Output:

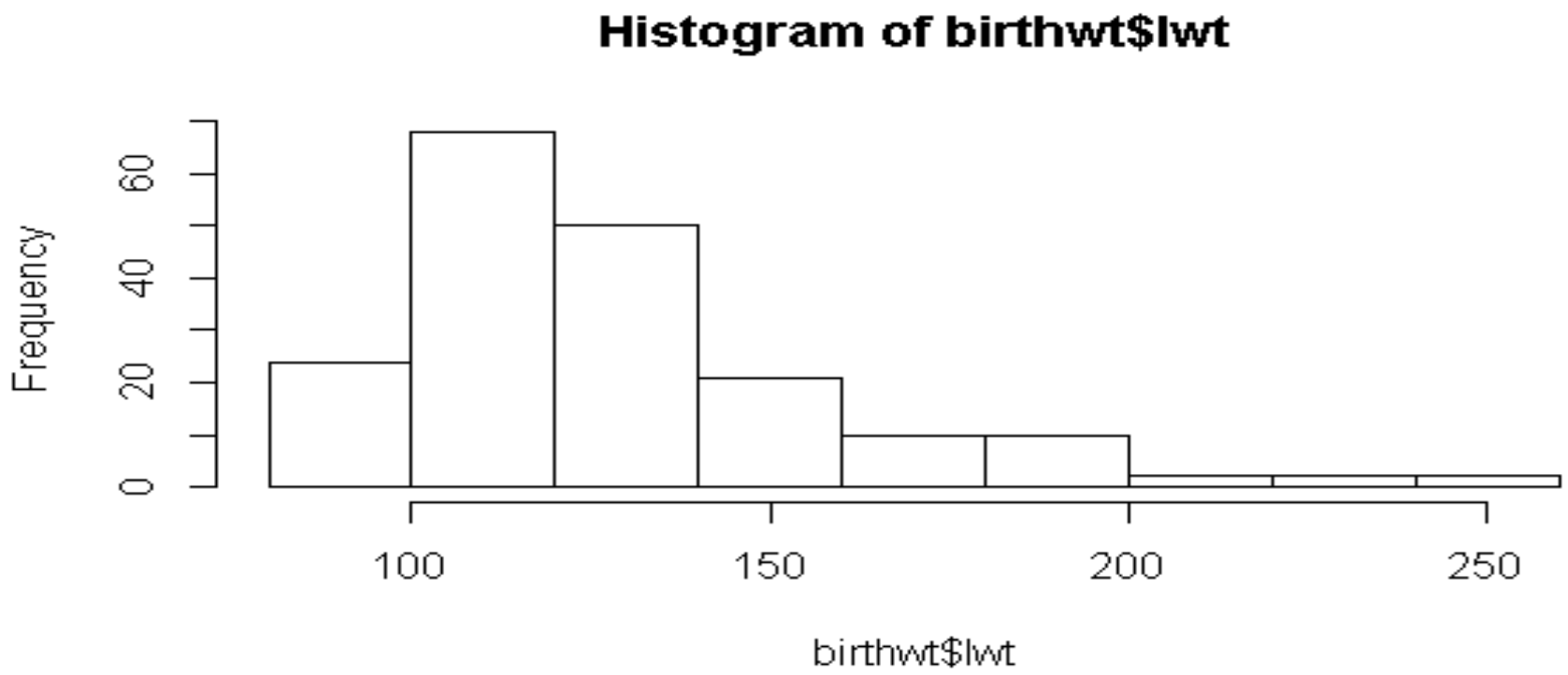


Fig 7: Non-Normal Distribution for birthwt$lwt using Histogram

TOOLS FOR INFERENTIAL STATISTICS

B. Hypothesis Testing

- Hypothesis can be defined as the description (i.e. a possible answer or a predictive statement) for a phenomenon while hypothesis tests are the procedures used for making coherent conclusions about some findings which are usually accompanied by a test of statistical significance (i.e. methods that are scientifically testable and measurable).

A. Types of Hypothesis

i. **Null Hypothesis (statistical hypothesis):** The null hypothesis is denoted by H_0 ('H-naught' or 'H-null'). Its interpretation implies that there is no relationship/difference between the existing variables of the population under investigation.

ii. **Alternative hypothesis (research hypothesis):** The alternative hypothesis is denoted by H_1 ('H-one' or H_a 'H-a'). Its interpretation means that a statement (finding) about two variables/groups under investigation is expected to be true.

- For example: Hypothesis testing example for the outcome of Bio-fertilizer 'x' on plant growth. We assume that a researcher has a new formulation Bio-fertilizer and wants to test the outcome on plant growth. According to Table I below, at the end of the experiment, we will either accept H_1 or H_0.

Table I: Experimental differences between Alternative Hypothesis (H_1) and Null Hypothesis (H_0)

Alternative Hypothesis	Null Hypothesis
It is the opposite of the null hypothesis.	It is the opposite of the alternative hypothesis.
A researcher wants to prove that the hypothesis is true.	A researcher wants to disprove or nullify that the hypothesis is true.
A researcher applies Bio-fertilizer 'x' and predicts that it improves plant growth.	A researcher applies Bio-fertilizer 'x' and predicts that it does not improve plant growth in any way.
This is for H_1: H_1 means that the application of Bio-fertilizer 'x' increases plant growth.	This is for H_0: H_0 means that the application of Bio-fertilizer 'x' does not increase plant growth.
The prediction states that there exist a statistical significance or relationship between the variables under investigation.	The prediction states that there does not exist a statistical significance or relationship between the variables under investigation.
If H_1 is accepted, it proves that the researcher's prediction is true.	If H_0 is accepted, it proves that the researcher's prediction needs to be re-examined.
Independent variable is Bio-fertilizer 'x'.	Independent variable is Bio-fertilizer 'x'.
Dependent variable is the plant. The plant is affected by the application of the independent variable 'x' which results into an increase in plant growth, increase in number of leaves, and increase in number of fruits. This means that the plant is dependent on Bio-fertilizer 'x'.	Dependent variable is the plant. The plant is not affected by the application of the independent variable 'x'. Hence, no result is expected. No increase in plant growth, no increase in number of leaves, and no increase in number of fruits. This means that the plant is not dependent on Bio-fertilizer 'x'.
The conclusion is that the independent variable can affect the dependent variable. Hence, H_1 predicts that there is a relationship between variable 'x' and the plant.	The conclusion is that the independent variable cannot affect the dependent variable Hence, H_0 predicts that there is no a relationship between variable 'x' and the plant.

B. Types of Errors

- When testing for statistical significance, there are two (2) types of errors that could occur during analysis. They are;

i. **Type I Error (false positive):** The type I error (denoted by α) occurs when H_0 is rejected because there isn't a true conclusion. Thus, it occurs when the null hypothesis is rejected instead of being accepted/retained. The probability of it occurring has a threshold that is set at 0.05 (i.e. the significance level).

ii. **Type II Error (false negative):** The type II error (denoted by β) occurs if we fail to reject H_0 when there is a true conclusion to prove the investigation. Thereby yielding into a false negative result.

- The difference between these two errors is that; type I error is falsely detects an outcome that is not existing, while a type II error is the failure to detect an outcome that exist.

C. *Level of Significance*

- Level of significance involves confidence level, p-value, and the one-tailed and two-tailed test. It is the probability of rejecting a null hypothesis by a conforming test when it is true and it is also denoted by α, i.e. P (type I error) = α.

i. **Confidence level (CI):** CI gives a predictable range of values where there is an undetermined population parameter calculated from a determined population sample. For example; An example from one of our projects, where we calculated values for the CI of a Mean (M) in a graph. Results are shown in Figure 10 below;

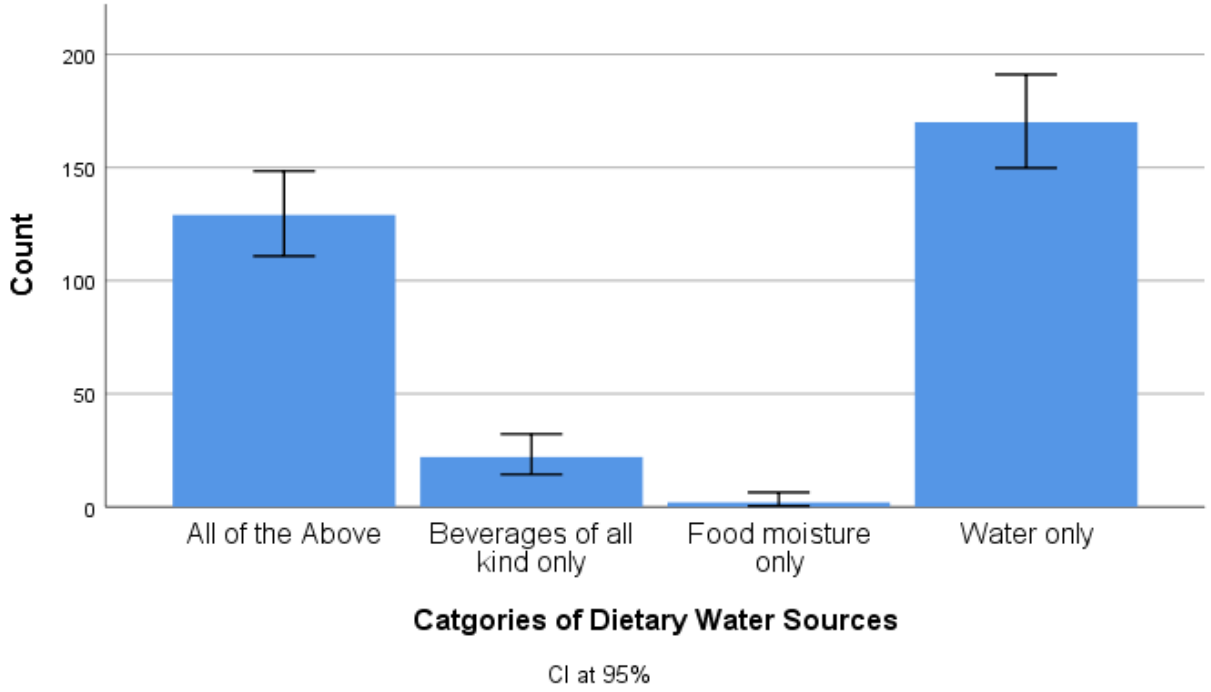


Fig 8: The participant's sources of TWI/dietary water

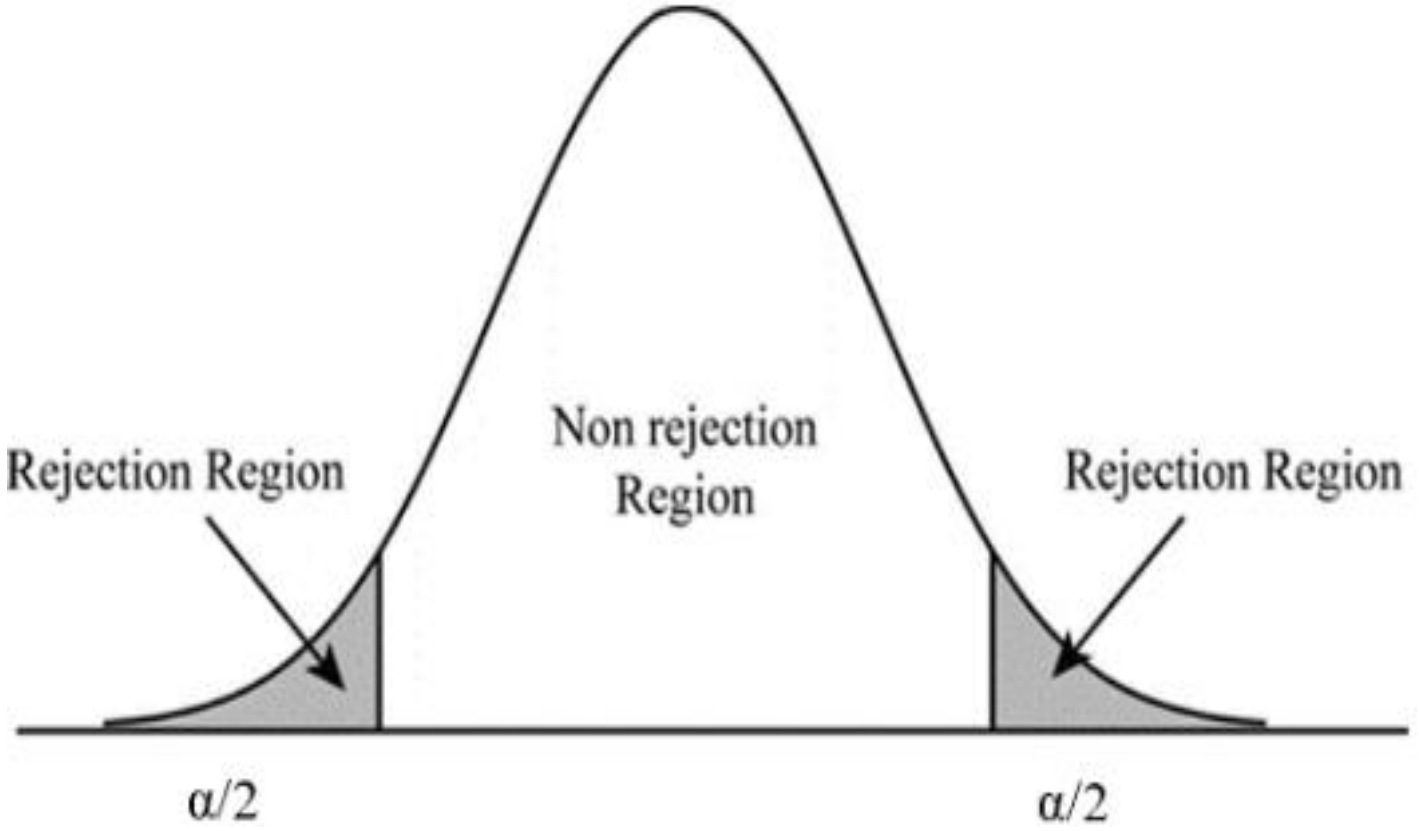
As shown in Figure 8, at 95% CI, there is no overlapping between participants who said only water is the source of dietary water and others who choose the rest of the category. Therefore, we conclude that a significant portion of the participants do not know the correct sources of dietary water.

ii. **P-value:** The p-value is defined as the probability that an event is likely to occur if H_0 is true. It is also known as the calculated probability. Probability is the measure of the likelihood that an event will occur (i.e. the possibility of an outcome) and it is usually represented as a number; 0 or 1 (where 0 represents uncertainty/impossibility and 1 represents certainty/possibility). P-value is a resulting number between 0 and 1 which can be used to interpret results where we decide if we want to reject/retain H_0. Therefore, to reject/retain the H_0, the following rejection rule should be considered;

- If p-value $\leq$ level of significance, then the null hypothesis (H_0) is being rejected.
- If p-value > level of significance, then the null hypothesis (H_0) is acceptable or not rejected.

iii. **One-tailed and two-tailed test:** The rejection region for two-tailed test is shown in Figure 9 below, while the rejection region for one-tailed test states that in the left-tailed test, the rejection region is shaded on the left side while in the right-tailed test, the rejection region is shaded on the right side

Fig 9: The rejection region for two-tailed test



iv. **Standard Error of Mean (SEM):** Standard error measures the accuracy a sample population represents the Mean (M) value of an entire population; the sample mean is called Standard Error of the Mean (SEM).

- For example: Let us explain better what SEM and SD are using illustrations from one of our projects; We calculated the Mean (M) value and SD value to check if the entire population knows what European Food Safety Authority (EFSA) and World Health Organization (WHO) says about the sources of dietary water and what they personally think it should be. The findings of this sample are best described by two (2) parameters; Mean (M) and SD. In Table II below, the Mean (M) of both results has no significant difference (i.e. the Mean (M) of "What do you think TWI should come from" is not significantly higher than "What does EFSA and WHO says TWI should come from") and indicates that both groups do not know the actual recommendation by EFSA and WHO.

Table II Mean and Standard Deviation values for TWI

	N	Mean	SD	SEM
What do you think TWI should come from	323	2.66	1.443	0.080
What does EFSA and WHO says TWI should come from	323	2.62	1.449	0.081

Table II above shows that the mean value for what the participants think TWI should come from is 0.04 more than the **Mean (M)** value of what the participants thinks EFSA and WHO recommends TWI should come from. Therefore, there is no significant difference, and we conclude that the population do not know the source of TWI.

CONCLUSION

- This paper has revealed that the most importance of biostatistics to biomedicine is the elucidation of raw data into knowledge. This paper may reveal the basic concept of biostatistics to researchers who are a novice in the use of R programming for statistical analysis.
- According to the methods used for statistical analysis in this paper, conclusion is drawn that; descriptive statistics is most useful in describing the association between variables from a population sample.
- Descriptive statistics provides a summary of the population in the form of **Correlation coefficient, Mean (M), Mode (Mo),** and **Median (Me)** while inferential statistics uses a random sample from the population to describe and make inferences about the whole population using **Null (H_0)** and **Alternative hypothesis (H_1), Confidence level** (**CI), Standard Deviation (SD), Standard Error of Mean (SEM)**, **Normal** and **Non-normal distribution** graphs and **P-value**.
- Inference statistics is valuable when it is not possible to examine each variable in the whole population

THANK YOU!

QUESTIONS?