INTRODUCTION TO BIOSTATISTICS

Dr Bornwell Sikateyo, PhD

Learning objectives

At the end of this lecture, you should be able to:

- Define statistics and biostatistics
- Define variable and random variable
- Differentiate different types of variables
- Distinguish between descriptive and inferential statistics
- Calculate means, medians, range, and standard deviations which are measures of central tendency and variability

Introduction to Biostatistics:

- Definition of Statistics

- Collection, organization, presentation, analysis, and interpretation of numerical data to assist decision-making

- The application of statistics to the public health (Biomedical) field is known as Biostatistics

Uses of Statistics

- Statistics presents a rigorous scientific method for gaining insight into data.
- For example, suppose we measure the weight of 100 patients in a study.
- With so many measurements, simply looking at the data fails to provide an informative account.
- However, statistics can give an instant overall picture of data based on graphical presentation or numerical summarization irrespective to the number of data points.

Uses of statistics

- Besides data summarization, another important task of statistics is to make inference and predict relations of variables.

Variables

- Variable - Any characteristic that can be measured or characterized

- Random variable - if each outcome of the variable can be determined by chance

Why Statistics?

Two Purposes

1. Descriptive

- Finding ways to summarize the important characteristics of a dataset

2. Inferential

- How (and when) to generalize from a sample dataset to the larger population

Descriptive Statistics

Statistics which provide graphical and numerical ways to organize, summarize, and characterize a dataset.

Inferential Statistics

Inferential statistics helps us make predictions or inferences about the population from the results of the sample.

Population: Is the totality of the observations of which we are concerned.

Sample: Is part or a subset of a population.

.

Inferential Statistics

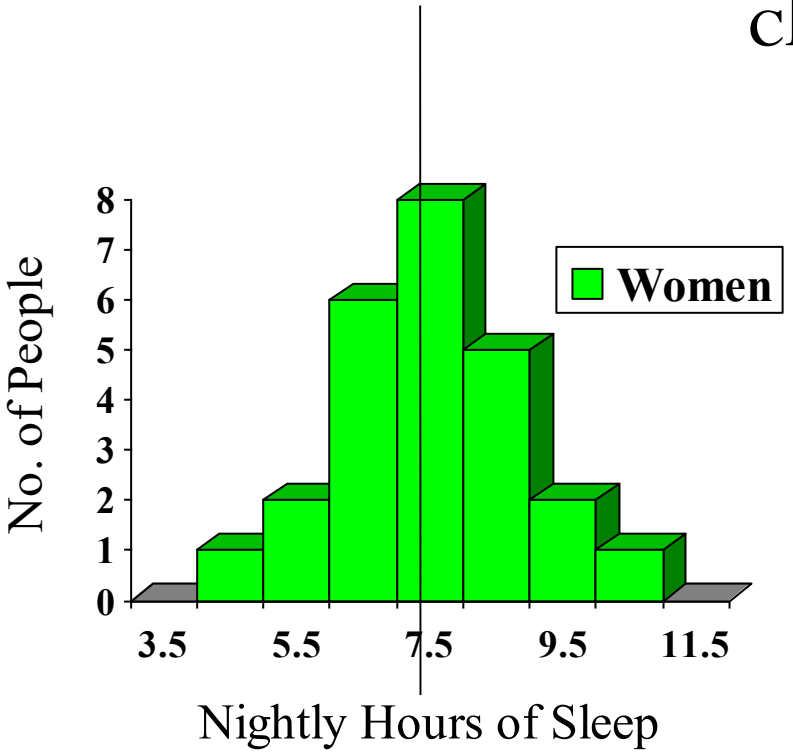
Population:

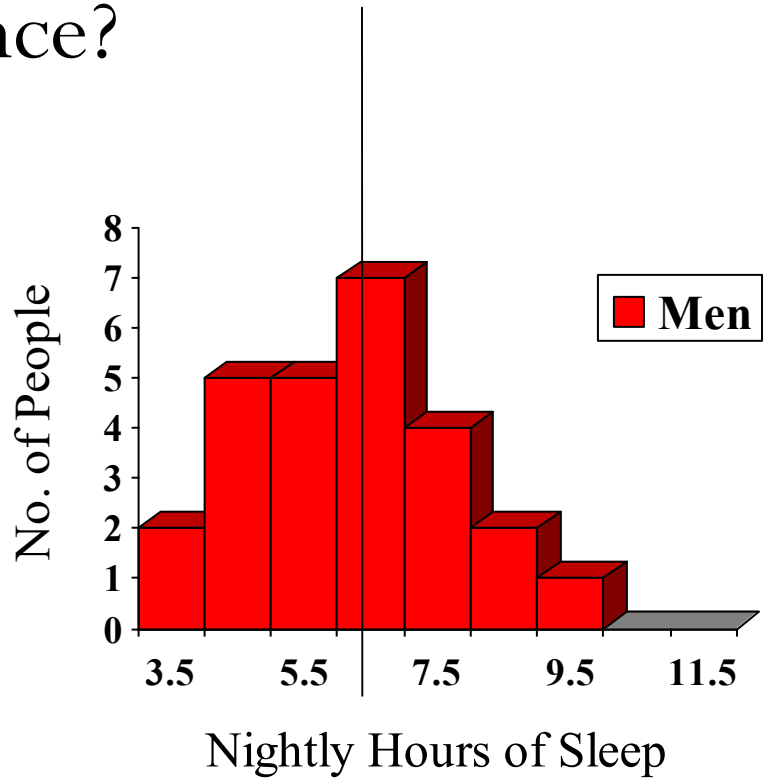
The set of all individuals of interest (e.g. all women, all college students)

Sample:

A subset of individuals selected from the population from whom data is collected

Are the sample differences simply due to chance?





Inferential Statistics

- Inferential statistics can tell us whether or not our results are likely to be due to chance alone

Some important terms

Parameter:

A characteristic of the population. Denoted with Greek letters such as μ or σ.

Statistic:

A characteristic of a sample. Denoted with English letters such as X or S.

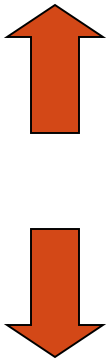
Sampling Error:

Describes the amount of error that exists between a sample statistic and the corresponding population parameter.

Important Point of Clarification

Statisticians ask: Was this observed "effect" caused by (lumpy) chance alone?

Inferential statistics separates



Random Causes:

Fluctuations of chance

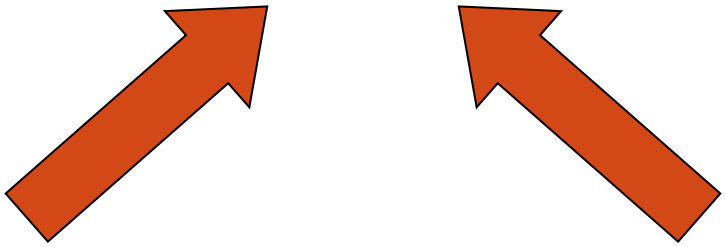
Non-random causes:

True differences in the population

Bias in the design of the study

A statistically significant result doesn't mean the results have to be "true". Just that they are non-random.

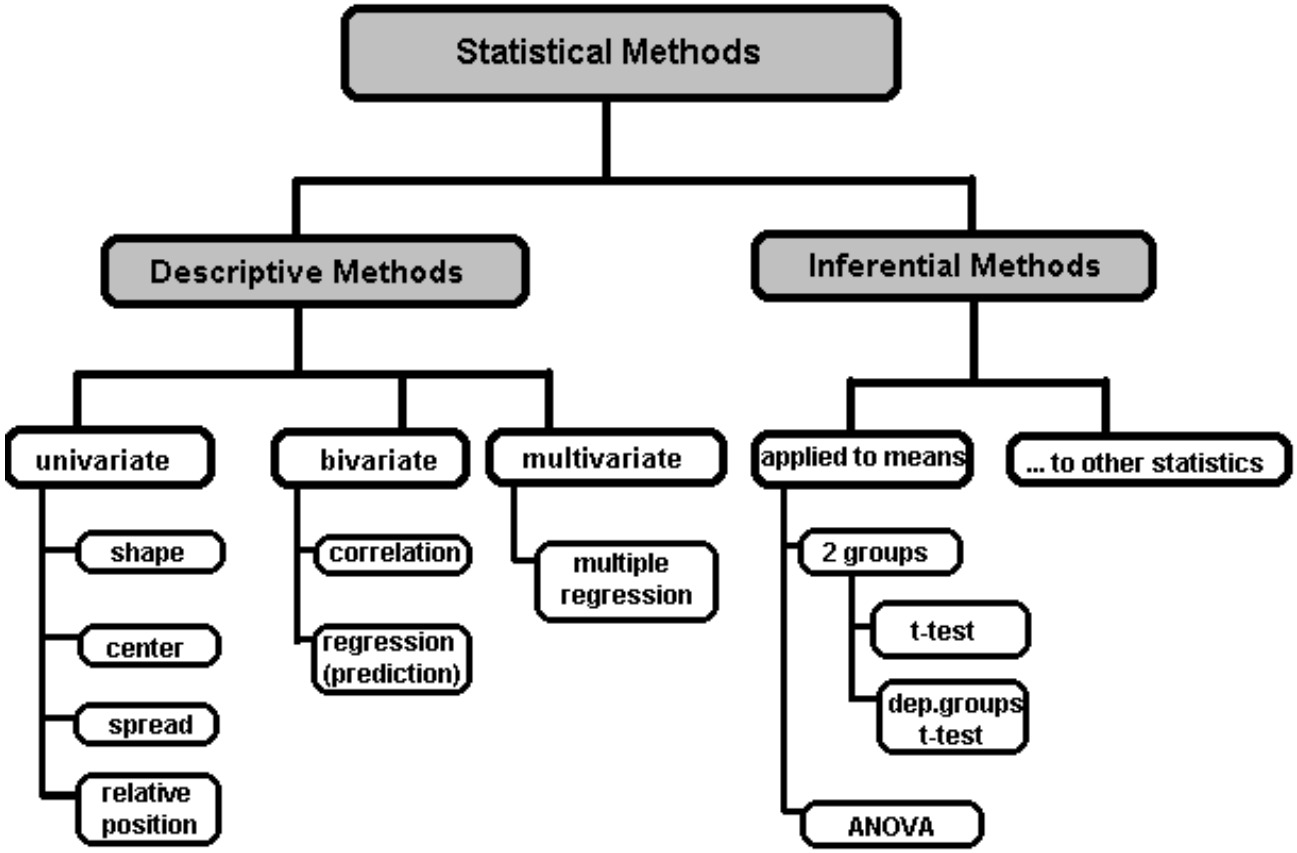
Inferential Statistics



Descriptive Statistics

Probability Theory

A Taxonomy of Statistics



Types of variables

- Continuous- numerical data measured on an unbroken scale
  - Age: . . .18, 19, 20, 21, 22, 23 . . .
  - Weight: . . .150, 151, 152, 153 . . .

- Categorical – data that are divided into distinct groups
  - Age Categories: 18-24, 25-44, 45-64, 65+
  - Gender: Male, Female

Types of Variables

Predictor variable:

The antecedent conditions that are going to be used to predict the outcome of interest. If an experimental study, then called an "independent variable".

Outcome variable:

The variable you want to be able to predict. If an experimental study, then called a "dependent variable".

Types of categorical variables

- **Ordinal-** any categorical variable with some intrinsic order or numeric value but not magnitude
  - Scales: On a scale of 1 to 5, how much do you like this lecture?

- **Nominal-** a categorical variable *without any* intrinsic order or magnitude
  - Place of residence: which province in Zambia do you live in?

- **Dichotomous-** binary variable is a categorical variable that has only 2 levels or categories
  - Yes/No question: did you eat Nshima this week?

Discrete data

- Ordering and magnitude important
- Numbers represent actual measurable quantities
- Restricted to integers or counts
- e.g. number of times a woman gives birth

Measures of central tendency

- The most widely investigated part of a data set is its centre because data usually tends to cluster at the center.

Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are mean, median, mode, geometric mean etc.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is $(20+30+40)/3 = 30$.

Notation: Let $x_1, x_2, \ldots x_n$ are n observations of a variable x. Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Mean

Or mean = $\frac{\text{sum of value}}{\text{\# of observations}}$

Example: age (yrs) of children (8, 11, 10, 8, 9)

Mean: $\frac{\text{sum}}{\text{\# obs.}} = \frac{8+8+9+10+11}{5} = \frac{46}{5} =$ 9.2 yrs

Mean

$$\text{Mean} = \frac{\text{sum of value}}{\text{\# of observations}}$$

Example: age (yrs) of individuals (8, 56, 10, 8, 9)

$$\text{Mean: } \frac{\text{sum}}{\text{\# obs.}} = \frac{8+8+9+10+56}{5} = \frac{91}{5} = 18.2 \text{ yrs}$$

- Adding an outlier (age = 56) changes the mean dramatically!

Median

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median.

For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Median

Median = Middle value

Example: age (yrs) of children (8, 11, 10, 8, 9)

Median: -Order data: 8,8,9,10, 11

-Pick the middle value

Here it is the 3rd: 9 years

Mean or Median

Strength of median: The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The median of these four observations is (30+40)/2 =35. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Thus **weakness of mean**: does not work well for skewed data and is not robust to outliers.

Weakness of median: It only relies on the central values and ignores all the other data.

Mode

- **Mode**: The value that is observed most frequently.

Problems with mode:

- The mode is undefined for sequences in which no observation is repeated.
- At times it might be nowhere near the centre of a data set.
- Sometimes there is more than one mode.

Mode

Observation that occurs most frequently

9 12 15 15 15 16 16 20 26

Observation	Number of occurrences
9	1
12	1
15	3
16	2
20	1
26	1

Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range*, *variance*, *standard deviation*, *interquartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It's a crude measure of variability.

Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2, \ldots x_n$ is

$$S^2 = \frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$

Variance of 5, 7, 3? Mean is $(5+7+3)/3 = 5$ and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation (s) : Square root of the variance. The standard deviation of the above example is 2.

Standard Deviation (s)

Example 2- Calculate the standard deviation for the following set of numbers:

9 12 15 15 15 16 16 20 26

$$\text{Standard deviation} = \sqrt{\text{Sum} \frac{(x_i - \bar{x})^2}{N - 1}}$$

Mean = **16** N = **9**

$[9-16]^2 + [12-16]^2 + [15-16]^2 + [15-16]^2 + [15-16]^2 + [16-16]^2 + [16-16]^2 + [20-16]^2 + [26-16]^2 = 184$

$SD = \sqrt{\frac{184}{8}} =$ **4.79**

Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the $\mathbf{(n+1)/4)q^{th}}$ observation of the data, where q is the desired quartile and n is the number of observations of data.

The **first quartile (Q1)** is the first 25% of the data. The **second quartile (Q2)** is between the 25^{th} and 50^{th} percentage points in the data. The upper bound of Q2 is the median. The **third quartile (Q3)** is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and **Q3** is the median of the second half of the ordered observations.

Methods of Variability Measurement

In the following example Q1= ((15+1)/4)1 =4th observation of the data. The 4th observation is 11. So Q1 of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1 Q2 Q3

The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Inter-quartile Range: Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 11=50. The middle half of the ordered data lie between 40 and 61.

Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

Coefficient of Variation = $\frac{\sigma}{\bar{x}} \times 100$

Shape of Data

- Shape of data is measured by
  - Skewness
  - Kurtosis

Skewness

- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

Kurtosis

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$