# BASIC CONCEPTS OF PROBABILITY

GABRIEL A. OKYERE (PhD)

KNUST

*goasare17@gmail.com*

January 20, 2017

# DEFINITIONS

## Statistics

Statistics is the science of collecting, organizing and interpreting numerical information or data.

**OR**

Statistics, in short, is the study of data. It includes descriptive statistics (the study of methods and tools for collecting data, and mathematical models to describe and interpret data) and inferential statistics (the systems and techniques for making probability-based decisions and accurate predictions.

## Inferential Statistics

Statistical Inference is the process of forming conclusions about the unknown parameters of a population by computing statistics from the individuals in a sample.

# IVPPSS

It is important that we understand the difference between **population** and **sample**, **parameter** and **statistic**, before we can understand and appreciate the process of making statistical inferences. Before identifying these items we must also identify the **individual** and **variable(s) of interest.** These six items must be explicitly identified at the beginning of any statistical analysis for that analysis to be conducted properly.

# Definition of **IVPPSS**

## Individual

An **individual** is one of the items examined by the researcher. An individual is not necessarily a person.

## Variable

A **variable** is the characteristic of interest about each individual.

## Population

A **population** is the collection of all individuals of interest.

## Parameter

A **parameter** is a summary of all individuals in the population. It is a number computed from the population.

# Definition of **IVPPSS** Cont'd

### Sample

A **sample** is a subset of the population examined by the researcher.

### Statistic

A **statistic** is a summary of the sample. It is a number computed from the sample.

# PERFORMING IVPPSS

## Steps for statistical analysis

- **First**, we determine what item we are actually going to look at; those are your individuals.
- **Second**, what are we going to record when we look at an individual, that is the variable.
- **Third**, the population is simply ALL of the individuals.
- **Fourth**, the parameter is the summary(e.g., mean or proportion) of the variable recorded from all the individuals in the population.
- **Fifth**, we realize that we cannot see all the individuals in the population so we examined a few (those few are the sample).
- **Finally**, the summary of the individuals of the sample is the statistic. The statistic has to be the same summary of the sample as the parameter was of the population

# EXAMPLE

## Example 1

My dad owns 60 acres of timber (mostly Oak, Walnut and Poplar) in Iowa. He wants to measure the mean-diameter-breast-height (DBH) of the oak trees on his property. He measures the DBH of 75 randomly selected oak trees. Use this information to perform an IVPPSS.

# SOLUTION

## Solution to Example 1

- Individual = an oak tree
- Variable = Diameter-breast-height (DBH)
- Population = All oak trees on Dad's property
- Parameter = mean DBH of all oak trees on Dad's property.
- Sample = 75 oak trees Dad measured
- Statistic = mean DBH of the 75 oak trees that Dad measured.

# VARIABLES

### Definition

A **variable** is the characteristic about each individual. The variable is the information that the researcher records about each individual. Note that in most "real life" studies the researcher will be interested in more than one variable.

Studies with one variable are called **univariate** studies, studies with two variables are **bivariate** studies, and studies with more than two variables are called **multivariate** studies.

# TYPES OF VARIABLES

There are two main groups of variable types - **quantitative** and **qualitative** variables.

## QUANTITATIVE VARIABLE

Quantitative variables are variables with numerical values for which it makes sense to do arithmetic operations (like adding or averaging).

## QUALITATIVE VARIABLE

Qualitative variables are variables that record to which group or category an individual belongs. Synonyms for qualitative are categorical or attribute.

Within each main type of variable are two subgroups.

# QUANTITATIVE VARIABLES

## TYPES OF QUANTITATIVE VARIABLES

The two types of quantitative variables are **continuous** and **discrete** variables.

1. Continuous variables are quantitative variables that have uncountable number of values. In other words, a potential value DOES exist between every pair of values of a continuous variable.
2. Discrete variables are quantitative variables that have countable number of values. Stated differently, a potential value DOES NOT exist between every pair of values of a discrete variable. Typically, but not always, discrete variables are counts of numbers.

# QUALITATIVE VARIABLES

## TYPES OF QUALITATIVE VARIABLES

The two types of qualitative variables are **ordinal** and **nominal**.

1. Ordinal variables are qualitative variables where a **natural order or ranking** exists among the categories.
2. Nominal variables are qualitative variables where a **NO order or ranking** exists among the categories.

# ORDINAL AND NORMINAL EXPLAINED

## Ordinal And Nominal

Ordinal and nominal variables are easily distinguished by determining if the order of the categories matters. For example, suppose that a researcher recorded a subjective measure of condition (i.e., poor, average, excellent) and the species of each duck. Order matters with th condition variable - i.e., the condition improves from the first (poor) to the last category (excellent) - and some re-orderings of the categories would not make sense, i.e., average, poor, excellent does not make sense. Thus, condition is an ordinal variable.

In contrast, species (eg, mallard, redhead, canvasback, and wood duck) is nominal because there is no inherent order among the categories (i.e., any reordering of the categories also "makes sense")

# UNIVARIATE EXPLORATORY DATA ANALYSIS

Once data has been collected, it is important to explore the distribution of the values of each variable. The first step in the statistical analysis is called Exploratory Data Analysis (EDA). The goal at this point is to develop a "feel" for the data, to identify what types of values each variable assumes and to determine if there are any 'issues' with the data.

# QUANTITATIVE UNIVARIATE EDA

## EXPLORING QUANTITATIVE DATA

A univariate EDA for quantitative data is concerned with describing the distribution of the values of a variable. Specifically, for each quantitative variable, the distribution is described by four specific attribution:

1. the **shape** of the distribution.
2. the presence of **outliers**
3. the measure of **location** of the distribution, and
4. the measures of **dispersion** or spread of the distribution.

# THE SHAPE OF THE DISTRIBUTION

In identifying the shape of a distribution, graphs are the best tools to use. Graphs also point out the presence of outliers and gives a general **depiction** for the center and dispersion of the data. Common graphs for quantitative exploratory data analysis are:

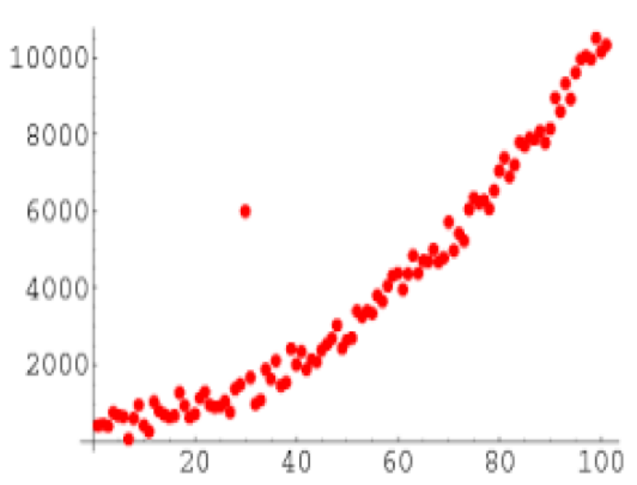
1. Histograms
2. Stem & Leaf Plot
3. Box Plot
4. Dot Plot

# THE PRESENCE OF OUTLIERS

An **outlier** is an individual in a sample whose value is widely separated from the main cluster of values in the sample.
In general, a group of more than two individuals are not considered as outliers even if they are separated from the main cluster of individuals.
Usually, outliers are removed so as not to influence the definition of the shape of the distribution but note that, not all outliers warrant removal from the sample.
In any case, it is important to plot the distribution of the data to determine if any outliers are present or not.

# THE PRESENCE OF OUTLIERS CONT'D

## A typical example showing an outlier



# THE MEASURE OF LOCATION OF THE DISTRIBUTION

The measures of location of the distribution can be categorized into measures of central tendency and measures of non-central tendency.

A **measure of central tendency** has the propensity to determine the central location or middle of the data. There are three common ways to measure the center of a distribution: **the mode, the median and the mean**.

A measure of non-central tendency determines the a specific location 'not necessarily the centre of a given data. These locations are commonly referred to as the **quantiles**. Examples of quantiles are percentiles and quartiles.

# THE MEASURE OF DISPERSION OF THE DISTRIBUTION

The degree to which a numerical data tend to spread about an average is called the dispersion of the data. Common methods for measuring the spread of a data are:

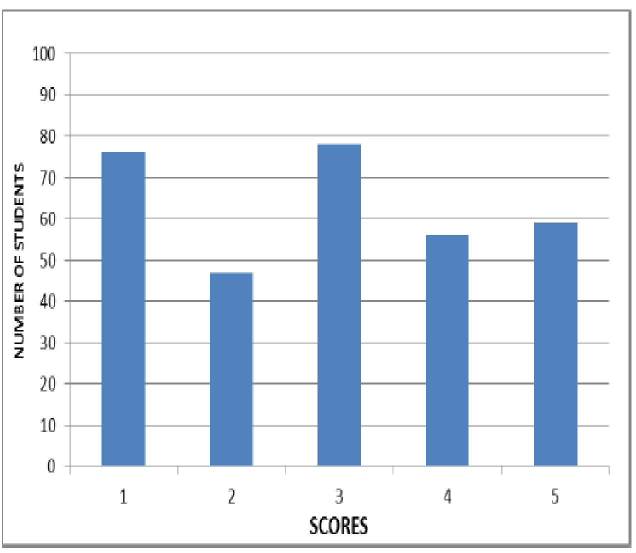
1. **The range**: It is the difference between the maximum and the minimum value in the data set. The range is never used on its own as a measure of dispersion.
2. **Inter-quartile range (IQR)**: It is the difference between the third and first quartiles
3. **The standard deviation**: It is 'essentially' the average deviation or difference of each individual from the mean.

# QUALITATIVE UNIVARIATE EDA

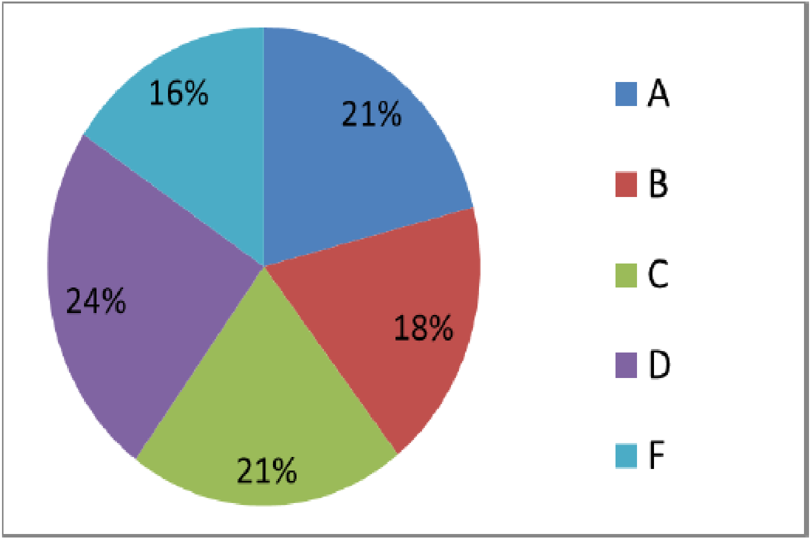
Interpreting summaries of a single categorical variable is more intuitive and less defined than that of quantitative data. Specifically one does not describe shape, location, outliers and dispersion for qualitative data. The two methods used are:

1. **Bar Chart**: It is used to display the frequency of individuals in the categories of a categorical variable. Bar plots have frequency of individuals on the y-axis and category labels plotted on the x-axis.
2. **Pie Chart**: It is a circular depiction of data where the area of the whole pie represents 100 percent of the data and slices of the pie represent a percentage breakdown of the sublevels.

# A typical bar chart showing the scores obtained by a group of students on a test



# A typical pie chart showing the percentages of students and their grades in statistics



# Trial Questions

## Questions

1. I have a friend who wants to start a (fishing) bait store on the West end of Ashland. He wants to determine what proportion of Ashland residents who currently use the East end bait store would use a store in the West end if one existed. He sends out 5000 questionnaires and receives 2378 back from patrons of the East end store. Use this information to perform an IVPPSS.
2. I'm interested in developing a model to predict how many points an NBA starting basketball player scores. Therefore, I want to determine the relationship between points scored and heights, speed(in the 40-yard dash), points and minutes played. To identify this relationship I gather these data from NBA starting basketball players. Use this information to perform an IVPPSS.

# Trial Questions Cont'd

## Try These Questions

1. You Might Be Interested To Know(YMBITK), the average level of mercury in newly-hatched goslings in the upper Midwest(MI, MN, ND, SD, WI). You obtained 20 goslings from resource agencies in each state. Use this information to perform an IVPPSS.
2. YMBITK, the proportion of NC students that think NC can become "the nation's leading environmental liberal arts college" in the next decade. You polled 124 students. Use this information to perform an IVPPSS.

# The End