Basic Statistics –Data Types & Sampling Techniques

Statistics - The science of collecting, organizing, describing, and interpreting data or information. In the study of statistics, it is important to be familiar with a variety of terms.

Population vs. a Sample - Data that is gathered can come from an entire population or by sampling that population.

Population – The complete set of people or things being studied *(i.e. All students at a certain high school)*

Parameters→ Numbers that are used to describe a population.

Sample – A subset of the population from which the raw data are actually obtained. *(i.e. polling 10% of students from every grade at a specific high school)* Sampling techniques are often utilized if it is not feasible to gather the entire population of data.

Statistics→ Numbers that are used to describe a sample.

Sampling Strategies - Often, it is not feasible to gather data from an entire population and sampling is used instead. There are a variety of sampling techniques that can be used to gather information. *The ultimate goal would be for the sample to be representative of the actual population.* Certain sampling techniques are more likely to achieve this goal. Studies conducted by procedures or sampling techniques that either over or underestimate a population's actual value are said to be **biased**.

Simple Random Sampling – A sample chosen by a method in which each collection of the population items is equally likely to make up the sample *(i.e. numbers in a lottery).* ***This method is often the most basic and best sampling method.***

Convenience Sampling - A sample that is not collected by a well-defined random method. Certain sample groups might be chosen over others because of the ease of access to them. *This method is not acceptable when it is possible that a systematic difference exists between the sample and the population. (i.e. Asking the first 20 people who leave a pet store whether they like dogs.)*

Stratified Sampling – A sample chosen by first dividing the population into groups (or strata) and then selecting a random sample from each strata. The members or items in each strata are similar in some way. *(i.e. Strata = 10 high schools in a certain city. Sample – select 100 students from every high school).*

Cluster Sampling - A sample chosen by first diving the population into groups (or clusters) and then selecting *(i.e. Clusters = 10 high schools in a certain city. Sample – select 3 of the high schools and poll every student at those 3 schools). This method is useful if the population is large and spread out.*

Systematic Sampling - A sample chosen by ordering the population and then selecting samples using a specific frequency. *(i.e. testing the brakes on every 10th car that is produced from an assembly line*). *This method is often used on assembly lines for quality control.*

Voluntary Response Sampling – A sampling method where individuals volunteer to participate. ***This method is highly biased and never reliable!*** *(i.e. Soliciting individuals to complete a survey about whether they are happy with the performance of a product.)*

Types of Data - Information that has been gathered is collected into a **Data Set**. Data Sets can be organized into lists, tables, and/or graphs. The specific data that is gathered is often referred to as a **Variable** and represents some characteristic of the population being studied.

Variables - The specific data that is gathered is often referred to as a Variable and represents some characteristic of the population being studied.

Qualitative – Classify variables into *categories*. There is **not** an associated number. *(i.e gender)* Qualitative variables can be ordinal or nominal.

- *Ordinal* – Categories have a natural order *(i.e. letter grades – A, B, C, D, F).*
- *Nominal* – Categories have no natural ordering *(i.e. colors – red, blue, green, …).*

Quantitative – Assigns variables a number so that it is apparent how much or how many of something there is. *(i.e. the daily high temperature in °F of a certain city in July).* Quantitative variables can be discrete or continuous.

- *Discrete* – A variable where every possibility can be listed, yet the list can be infinite *(i.e. the list of integers).*
- *Continuous* – A variable that can take on any value within some interval. Each of the possibilities cannot be restricted to a list *(i.e. people's heights or weights).*

Types of Experiments - Experiments are often used to gather information in order to draw conclusions regarding a situation or set of circumstances. For example, doctors may use experiments and statistics to determine if a new drug is effective in the treatment of a certain disease.

Experimental Units (EU) - The individuals that are being studied *(people, animals, plants, etc.)*. If the units are people, they may also be referred to as **subjects**.

Outcome / Response - The variable being measured on each EU. *(i.e. the size of a tomato).*

Treatments - The procedures applied to each experimental unit. There must be at least 2 treatments. *(i.e. No fertilizer and Fertilizer A).* The purpose is to determine whether the treatment affects the outcome *(i.e. how large the tomatoes grow).*

Randomized Experiments - A study in which the investigator assigns the treatment(s) to the experimental units at random. *(i.e. For 3 tomato plants, an investigator assigns a) No fertilizer, b) Fertilizer A, and c) Fertilizer B.)*

Double Blind - An experiment in which neither the investigator nor the subjects know who has been assigned which treatment(s).

Observational Studies - A study in which the assignment to treatment groups ***is not*** made by the investigator. *(i.e. Studying smokers vs. non-smokers).*

Confounder – A variable that is related to both the treatment and the outcome. Confounders make it difficult to determine if differences in the outcomes are due to the treatments or not. *(i.e. Studying the effects of smoking (treatment) on liver disease (outcome). Alcohol consumption could be a confounder.)*