LECTURE 1: INTRODUCTION AND BASIC CONCEPTS

1. Introduction to statistics

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data[1]. In Statistics, the phenomena being studied should be quantified and therefore are reduced to numbers. Statistics is comprised of two main branches, a) descriptive statistics and b) inferential statistics.

a) "**Descriptive statistics** can help to provide a simple summary or overview of the data, thus allowing researchers to gain a better overall understanding of the data" (Mackey & Gass, 2005, pp. 250-251). This summary of the data is done through measures of frequency, measures of central tendency (i.e., the mode, median, and mean), and measures of spread (i.e., variance and standard deviation).

b) Inferential statistics refer to statistical tests that allow the researchers to infer the results obtained with a sample to the general population (hence the name inferential). Such tests are very important as they ensure the researcher that the difference found between tests or groups is not due to chance but rather to a real relationship (causal or correlational) between the variables being studied.

2. Important concepts in statistics

2.1. Data

The term data (plural for datum) refers to facts, information, and observations about an object (or objects), a phenomenon (or phenomena), a person (or a group of people). Data can take either a qualitative or a quantitative form.

Qualitative data refers to the type of descriptive information that does not make use of statistical procedures (Mackey & Susan, 2015, p. 215). Such data are often collected using observations, questionnaires, or interviews. Instead of quantification, the collection and analysis of qualitative data involves careful and detailed descriptions of objects, phenomena, or people's narratives, perceptions, and beliefs about objects or phenomena.

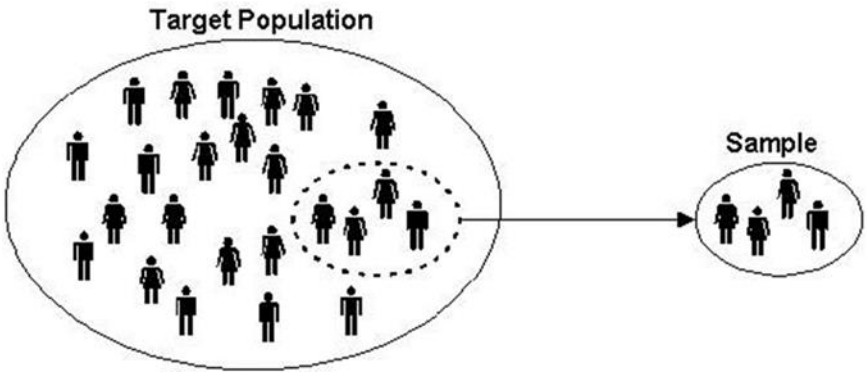
Quantitative data consists of information that is, in some way or the other, quantifiable. In other words, we can put quantitative data into numbers, figures, and graphs, and process it using statistical procedures. When

[1] Romijn, Jan-Willem (2014). "Philosophy of statistics". Stanford Encyclopedia of Philosophy. Retrieved from https://plato.stanford.edu/entries/statistics/

using quantitative analysis, we are usually interested in how much and how many there is/are of whatever we are interested in. The ultimate goal of quantitative data analysis is to generate facts that are generalizable.

2.2. A population vs. a sample

A **population** is the entire group that the researcher wants to draw conclusions about. A **sample** is the specific group that the researcher will collect data from. The size of the sample is always less than the total size of the population.



2.3. Variables

In research, variables refer to the population, objects, or phenomena that the researcher is trying to measure.

Dependent variable: The variable that is influenced by the independent variable.

Independent variable: The factor that causes the dependent variable to change.

Example: The influence of **English media** on ***EFL learners' listening comprehension.***

2.4. Hypotheses

Research hypotheses can be used to express what the researcher expects the results of the investigation to be. The hypotheses are based on observations or on what the literature suggests the answers might be (Mackey & Gass, 2005: 19).

a) Null hypothesis: (H_0) is a neutral statement used as a basis for testing. The null hypothesis states that there is no relationship between items under investigation. In advanced inferential statistical tests, the objective is often to reject the null hypothesis by providing a proof that there is a relationship between variables X and Y.

b) Non-directional hypothesis (two-tailed/ two-way hypothesis): In these types of hypotheses, a difference between variables under investigation is predicted, but the researcher is not interested in the direction of the difference/ relationship., but its direction is not specified.

Examples:

- Text-to-speech technologies will affect Algerian EFL learners' pronunciation awareness.
- There will be a relationship between family income and students' academic performance.

c) Directional hypothesis (one-tailed): In these types of hypotheses, the researcher indicates the predicted direction of the relationship between two or more variables.

Examples:

- Older learners will prefer speaking more than younger learners.
- Novice teachers will give less instructions in the target language than experienced teachers.

2.5. Causation vs. correlation

"Correlational research attempts to determine the relationship between or among variables; it does not determine causation." (Mackey & Gass, 2005, p. 284). On the other hand, causal relationships are unveiled by determining the causes and effects of a particular phenomenon. In research, there is a general consensus that correlation does not equal causation. In other words, just because there a relationship between two variables, it does not mean they cause each other. For example, high classroom attendance is often correlated with high academic achievement. However, it is not necessary that high classroom attendance is causing the high academic achievement (or the opposite). Perhaps, there is a third factor (e.g., willingness to study or proficiency) that is directly contributing to academic achievement.

2.6. Probability

In statistics, probability refers to the likelihood of an event taking place.[2] For example, the probability of getting heads or tails when tossing a coin is 50%. In statistical tests, the probability value (or p-value) measures the probability that an observed relationship/ difference could have occurred due to random chance. Low p-values ($p < 0.05$) indicate a great chance of real statistical significance (i.e., rejecting the null hypothesis); meanwhile, high p-values indicate a low chance of statistical significance (i.e., accepting the null hypothesis).

2.7. Bias

"Bias takes place when a statistical model/ test over or underestimates a relationship between two or more variables. There are two main types of bias, a) selection bias and b) confirmation bias. **Selection bias** happens when a researcher tries to generalize findings that were collected in a non-random way. **Confirmation bias**, on the other hand, takes place when the researcher starts the data collection/ analysis with a predetermined assumption." [2]

[2] Vickery, Rebecca. (2021, January 30). "8 Fundamental Statistical Concepts for Data Science". https://towardsdatascience.com/8-fundamental-statistical-concepts-for-data-science-9b4e8a0c6f1c