STATISTICS

INTRODUCTION

- Statistics can be defined as the practice of collecting, organizing, describing, and analyzing data to draw conclusions from the data to apply to a cause. Data must either be numeric in origin or transformed by researchers into numbers.
- There are two main branches of statistics: descriptive and inferential. **Descriptive statistics** is used to say something about a set of information that has been collected only. **Inferential statistics** is used to make predictions or comparisons about a larger group (a population) using information gathered about a small part of that population.

DESCRIPTIVE STATISTICS

- Descriptive statistics describe patterns and general trends in a data set. In most cases, descriptive statistics are used to examine or explore one variable at a time. However, the relationship between two variables can also be described as with correlation and regression.
- The first phase of data analysis involves the placing of some order on some sort of "chaos". Typically the data are reduced down to one or two descriptive summaries

FREQUENCY DISTRIBUTIONS

- These types of distributions are a way of displaying “chaos” of numbers in an organized manner so such questions can be answered easily. A frequency distribution is simply a table that, at minimum, displays how many times in a data set each response or "score" occurs.

MEASURES OF CENTRAL TENDENCY

- This type of measurements gives a single description of the average or "typical" score in a data distribution. These measures attempt to quantify what we mean when we think of as the "average" score in a data set. The concept is extremely important and we encounter it frequently in daily life.
- For example, we often want to know before purchasing a car its average distance per liter of petrol. Or before accepting a job, you might want to know what a typical salary is for people in that position so you will know whether or not you are going to be paid what you are worth.

MEASURES OF CENTRAL TENDENCY

- 1. **Median:** It is the middle number of a set of numbers arranged in numerical order. If the number of values in a set is even, then the median is the sum of the two middle values, divided by 2.
- 2. **Mode:** the mode is the most frequent value in a set. A set can have more than one mode; if it has two, it is said to be bimodal. The mode is useful when the members of a set are very different, for example the comparison of grades of a test (A, B, C, D, E).
- 3. **Mean:** the mean is the sum of all the values in a set, divided by the number of values. The mean of a whole population is usually denoted by μ, while the mean of a sample is usually denoted by $\bar{x}$.

MEASURES OF VARIABILITY

- The average score in a distribution is important in many research contexts. So too is another set of statistics that quantify how variable or how dispersed the scores in a set of data tend to be. Sometimes variability in scores is the central issue in a research question. Variability is a quantitative concept, so none of this applies to distributions of qualitative data.
- **1. Range:** the range is the difference between the largest and smallest values of a set, but is not very useful because it depends on the extreme values, which may be distorted.

MEASURES OF VARIABILITY

- **2. Variance:** the variance is a measure of how items are dispersed about their mean. The variance of a whole population is given by the equation: $\sigma^2 = \frac{\sum(x-\mu)^2}{N}$. The variance of a sample is calculated differently: $s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$
- **3. Standard deviation:** The standard deviation "σ" (or "s" for a sample) is the square root of the variance.
- **4. Relative variability:** The relative variability of a set is its standard deviation divided by its mean. The relative variability is useful for comparing several variances.

INFERENTIAL STATISTICS

- Inferential statistics are used to judge the meaning of data. Inferential statistics assess how likely it is that group differences or correlations would exist in the population rather than occurring only due to variables associated with the chosen sample.
- Two basic uses of inferential statistics are possible:
  - **Confidence intervals,** which is also referred to as Interval estimation.
  - **Hypothesis testing,** which is also referred to as Point Estimation.

THE HYPOTHESIS TESTING

- Often times we want to determine whether a claim is true or false. Such a claim is called a hypothesis. It is important to clear up some terms such as:

1. **Null hypothesis:** a specific hypothesis to be tested in an experiment. The null hypothesis is usually labeled Ho.

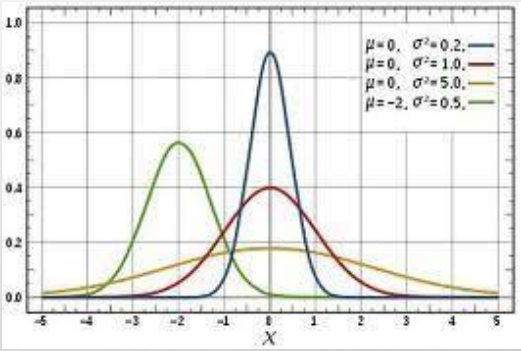
2. **Alternative hypothesis:** a hypothesis that is different from the null hypothesis, which we usually want to show that is true (thereby showing that the null hypothesis is false). The alternative hypothesis is usually labeled Ha.

THE HYPOTHESIS TESTING

- The null hypothesis is tested through the following procedure:
  - Determine the null hypothesis and an alternative hypothesis.
  - Pick an appropriate sample.
  - Use measurements from the sample to determine the likelihood of the null hypothesis.
- Other important concepts when studying hypothesis test are:
  - **Type 1 Error:** If the null hypothesis is true but the sample mean is such that the null hypothesis is rejected, a Type I error occurs. The probability that such an error will occur is the risk.
  - **Type 2 Error:** If the null hypothesis is false but the sample mean is such that the null hypothesis cannot be rejected, a Type II error occurs. The probability that such an error will occur is called the risk.

CONFIDENCE INTERVALS

- Interval estimation is used when we wish to be fairly certain that the true population value is contained within that interval. When we attach a probability statement to an estimated interval, we obtain a confidence interval.
- Confidence is defined as 1 - alpha (1 minus the significance level). Thus, when we construct a 95% confidence interval, we are saying that we are 95% certain that the true population mean is covered by the interval - consequently, of course, we have a 5% chance of being wrong.



- Any statistic that can be evaluated in a test of significance ("hypothesis testing") can be used in constructing a confidence interval.