INTRODUCTION TO STATISTICAL METHODS

I. Statistics and the Statistical Method

What is Statistics?

- Statistics is both an art and a science of designing studies and analyzing the information that those studies produce.
- Its ultimate goal is translating data into knowledge and understanding of the world around us.
- In short, statistics is the art and science of learning from data.
- The word "statistics" is used in 3 main ways:
- Common meaning: factual information involving numbers. A better word for this is data.
- Precise meaning: quantities which have been derived from sample data, e.g. the mean (or average) of a data set
- Common meaning: an academic subject which involves reasoning about statistical quantities

Why study Statistics?

- Statistics Helps Us Learn about the World
  - Explore a wide variety of everyday scenarios
  - Evaluate media reports about opinion surveys, medical research studies, the state of the economy, and environmental issues.
  - Make financial decisions
- Understanding of statistics is essential for making good decisions in an uncertain world.
- Most professional occupations today rely heavily on statistical methods.
  - In a competitive job market, understanding statistics provides an important advantage.
  - Court trials use probability and statistical inference in evaluating the quality of evidence.

Some important notes on the use of Statistics…

- **Business**
  - analyze results of marketing studies about new products, to help predict sales, and to measure employee performance.
- **Finance**
  - study stock returns and investment opportunities
- **Medicine**
  - evaluate whether new ways to treat disease are better than existing ways

Using Data to Answer Statistical Questions

- **Information gathering** is the heart of investigating answers to questions
- **Data**
  - information we gather with experiments and survey

Using Data to Answer Statistical Questions

- What data do we need to collect?
- Does a low-carbohydrate diet result in significant weight loss?
  - weight at the beginning of the study, weight at the end of the study, number of calories of food eaten per day, carbohydrate intake per day, body-mass index (BMI) at the start of the study, and gender.
- Are people more likely to stop at a Starbucks if they've seen a recent Starbucks TV commercial?
  - percentage of people who went to a Starbucks since the ad aired and analyze how it compares for those who saw the ad and those who did not see it

4-Step Process in Learning from Data

- Defining the problem
- Collecting the data
- Summarizing the data
- Analyzing data, interpreting the analyses, communicating the results

- **Scientific method**
  - Formulation of research goals, design of observational/experimental studies, collection of data, modelling/analysis and testing hypotheses

Statistics is a science…

- Statistical problem solving is an investigative process that involves four components:
  1. formulate a statistical question
  2. collect data,
  3. summarize the data, and
  4. analyze data and interpret results

Statistics: Three Main Areas

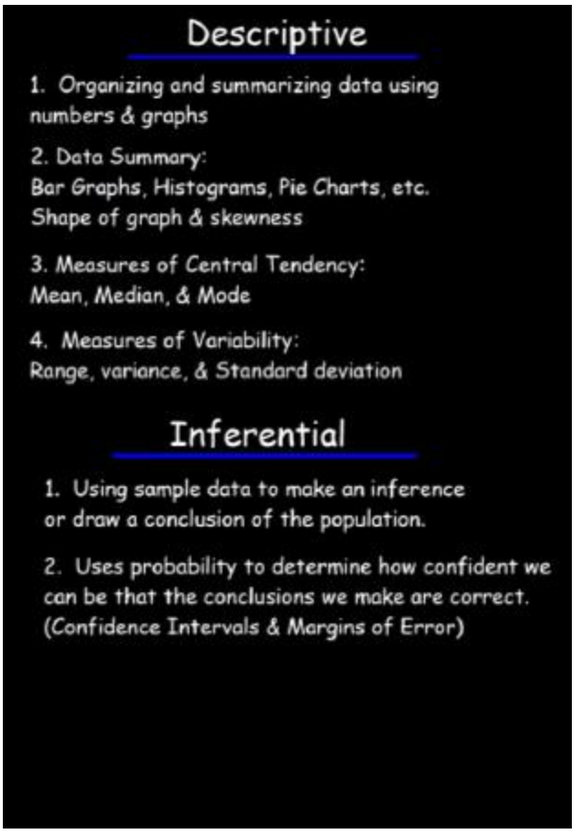
- **Descriptive statistics** – describing and summarizing data sets using pictures and statistical quantities
- **Inferential statistics –** analyzing data sets and drawing conclusions from them
- **Probability** – the study of chance events governed by rules (or laws)

Statistics: Data types

- In statistics it is vital to understand what types of data you are working with.
- There are three main types:
  - **Nominal** – categories that do not have a natural order, e.g. gender, eye color, types of building
  - **Ordinal** – categories which have a natural order but are not numerical, e.g. Likert scales (Strongly Disagree, Disagree, Neither Agree Nor Disagree, Agree, Strongly Agree)
  - **Scale/continuous** – numerical data ordered against a constant scale, e.g. date, temperature, length, weight, frequency.

Descriptive Statistics and Inferential Statistics

- **Descriptive statistics** refers to methods for summarizing the collected data (where the data constitutes either asample or a population).
  - The summaries usually consist of graphs and numbers such as averages and percentages.
- **Inferential statistics** refers to methods of making decisions or predictions about a population, based on data obtained from a sample of that population.
  - Use descriptive statistics to summarize the sample data and inferential statistics to make predictions about the population.



II. Sampling Techniques and Methods of Collecting Data

Sampling Techniques and Methods of Collecting Data

Sample and Population

- The population is the total set of subjects of interest.
- A sample is the subset of the population for whom we have (or plan to have) data, often randomly selected.

Sample Size

- Sample size refers to the number of individuals, elements, or observations included in a study or experiment.

Something to Think About…

- Think about the last time you made a decision based on a small sample.
- What are the potential implications of using a small sample size for making important decisions?

The potential implications of using a small sample size in this situation could be:

- Limited Representativeness
- Increased Sampling Error
- Biased Conclusions
- Unreliable Generalization
- Risk of Misinterpreting Trends
- Limited Confidence in Results

Importance of Sample Size in Statistical Analysis

- Sample size is a critical factor in statistical analysis, influencing the reliability and generalizability of study findings.

Importance of Sample Size in Statistical Analysis

- A larger sample size increases the likelihood of the sample accurately reflecting the characteristics of the population.
- This enhances the external validity of the study.

Importance of Sample Size in Statistical Analysis

- Larger sample sizes result in smaller margins of error and higher precision.
- This means that the estimate derived from the sample is more likely to be close to the true population parameter

Importance of Sample Size in Statistical Analysis

- Adequate sample sizes improve the statistical power of a study, increasing the likelihood of detecting a true effect if it exists.
- Low statistical power can lead to the failure to detect real differences or relationships.

Sampling Techniques

- Sampling techniques are methods used to select a subset of individuals or items from a larger population for the purpose of making inferences or generalizations about the entire population.

Random Sampling

- Random sampling involves selecting individuals or items from a population in a way that each member has an equal and independent chance of being chosen.
- This is often achieved through randomization methods.
- Simple random sampling refers to any sampling method that has the following properties:
  - The population consists of N objects.
  - The sample consists of n objects, and all possible samples of n objects are equally likely to occur.

Random Sampling

- **Simple Random Sampling**
  - Each individual in the population has an equal chance of being selected.
- **Systematic Random Sampling**
  - A random starting point is chosen, and every nth individual is included in the sample.
- **Stratified Random Sampling**
  - The population is divided into strata, and random samples are taken from each stratum.
- **Cluster Sampling**
  - Cluster sampling is a sampling technique in which the population is divided into clusters or groups, and then a random sample of clusters is selected for analysis.
  - With cluster sampling, every member of the population is assigned to one, and only one, group.
  - A sample of clusters is chosen, using a probability method (often simple random sampling). Only individuals within sampled clusters are surveyed.
  - Population divided into several “clusters,” each representative of the population
  - Simple random sample selected from each
  - The samples are combined into one

III. SAMPLINGTECHNIQUES AND METHODS OF COLLECTING DATA

Sample Size

- Understanding how to determine the appropriate sample size is crucial for ensuring the accuracy and reliability of research findings
- A sample size that is too small may not accurately represent the population, leading to biased results, while an excessively large sample

size may be unnecessarily costly and time-consuming
- the number of observations used for determining the estimations of a given population
- the number of completed responses your survey receives

Margin of Error
- the degree of error in results received from random sampling surveys
- a percentage that tells you how much you can expect your survey results to reflect the views of the overall population

Confidence Level
- a percentage that reveals how confident you can be that the population would select an answer within a certain range

Population Size
- total number of people in the group you are trying to study

Variability
- describes how far apart data points lie from each other and from the center of a distribution

Sampling Techniques
- method of selecting individual members or a subset of the population to make statistical inferences from them and estimate the characteristics of the whole population

Random Sampling
- each sample has an equal probability of being chosen

Simple
- researcher randomly selects a subset of participants from a population

Simple Random Sampling

Stratified
- researchers divide subjects into subgroups called strata based on characteristics that they share
- once divided, each subgroup is randomly sampled using another probability sampling method

Cluster
- you divide a population into clusters, such as districts or schools, and then randomly select some of these clusters as your sample

Systematic
- researchers select members of the population at a regular interval

Non-Random Sampling
- the sample selection is based on factors other than just random chance

Convenience
- units are selected for inclusion in the sample because they are the easiest for the researcher to access

Quota
- relies on the non-random selection of a predetermined number
- you first divide the population into strata and then recruit sample units until you reach your quota

Purposive
- select a specific group of individuals or units for analysis

Snowball
- new units are recruited by other units to form part of the sample

Data Collection
- process of gathering information from all the relevant sources to find a solution to the research problem

Quantitative
- on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures

Experiment
- Manipulate variables and measure their effects on others
- To test a causal relationship

Surveys
- Distribute a list of questions to a sample online, in person or over-the-phone
- To understand the general characteristics or opinions of a group of people

Qualitative
- on mathematical calculations using various formats like close-ended questions, correlation and regression methods, mean, median or mode measures

Interview/Focus Group
- Verbally ask participants open-ended questions in individual interviews or focus group discussions
- To gain an in-depth understanding of perceptions or opinions on a topic

Observation
- Measure or survey a sample without trying to affect them
- To understand something in its natural setting

Ethnography
- Join and participate in a community and record your observations and reflections
- To study the culture of a community or organization first-hand

Archival Research
- Access manuscripts, documents or records from libraries, depositories or the internet
- To understand current or historical events, conditions or practices

Secondary Data Collection
- Find existing datasets that have already been collected, from sources such as government agencies or research organizations
- To analyze data from populations that you can't access first-hand

IV. Graphic Displays of Basic Statistical Descriptions of Data

Introduction

- There are graphic displays of basic statistical descriptions. These include quantile plots, quantile–quantile plots, histograms, and scatter plots. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Outliers

- There are graphic displays of basic statistical descriptions. These include quantile plots, quantile–quantile plots, histograms, and scatter plots. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing. The first three of these show univariate distributions (i.e., data for one attribute), while scatter plots show bivariate distributions (i.e., involving two attributes).

Univariate descriptive statistics

- Univariate descriptive statistics focus on only one variable at a time. It's important to examine data from each variable separately using multiple measures of distribution, central tendency and spread. Programs like SPSS and Excel can be used to easily calculate these.

Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute (allowing the user to assess both the overall behavior and unusual occurrences). Second, it plots quantile information.

Quantile-Quantile Plot

- A quantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Histograms

- Histograms (or frequency histograms) are at least a century old and are widely used. "Histos" means pole or mast, and "gram" means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X.
- If X is nominal, such as automobile model or item type, then a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a bar chart.
- If X is numeric, the term histogram is preferred. The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width.
- Typically, the buckets are of equal width. For example, a price attribute with a value range of $1 to $200 (rounded up to the nearest dollar) can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on. For each subrange, a bar is drawn with a height that represents the total count of items observed within the subrange.

Scatter Plot

- A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.
- Figure 7 shows a scatter plot for the set of data in Table 1. The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.
- Two attributes, X, and Y, are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated).