WEEK – 1: 26/06/23

BASIC TERMS OF STATISTICS:

1. Population : All the members of a group about which you want to draw a conclusion.
2. Sample: The part of the population selected for analysis.
3. Parameter: A numerical measure that describes a characteristic of a population.
4. Statistic: A numerical measure that describes a characteristic of a sample.
5. Variable: A characteristic of an item or an individual that will be analyzed using statistics.

BRANCHES OF STATISTICS:

DESCRIPTIVE STATISTICS:

The branch of statistics that focuses on collecting, summarizing, and presenting a set of data.

Example: The average age of citizens who voted for the winning candidate in the last presidential election, the average length of all books about statistics, the variation in the weight of 100 boxes of cereal selected from a factory’s production line.

INFERENTIAL STATISTICS:

The branch of statistics that analyzes sample data to draw conclusions about a population.

Example: A survey that sampled 2,001 full- or part-time workers ages 50 to 70, conducted by the American Association of Retired Persons (AARP), discovered that 70% of those polled planned to work past the traditional mid60s retirement age. By using methods discussed in Section 6.4, this statistic could be used to draw conclusions about the population of all workers ages 50 to 70.

SOURCES OF DATA:

PUBLISHED SOURCES : Data available in print or in electronic form, including data found on Internet Web sites. Primary data sources are those published by the individual or group that collected the data. Secondary data sources are those compiled from primary sources.

Many U.S. federal agencies, including the Census Bureau, publish primary data sources that are available at the Web site www.fedstats.gov. Business news sections of daily newspapers commonly publish secondary source data compiled by business organizations and government agencies

EXPERIMENTS : T A process that studies the effect on a variable of varying the value(s) of another variable or variables, while keeping all other things equal. A typical experiment contains both a treatment group and a control group. The treatment group consists of those individuals or things that receive the treatment(s) being studied. The control group consists of those individuals or things that do not receive the treatment(s) being studied.

SURVEYS : A process that uses questionnaires or similar means to gather values for the responses from a set of participants.

SAMPLING CONCEPTS:

SAMPLING : The process by which members of a population are selected for a sample

PROBABILITY SMAPLING : T A sampling process that takes into consideration the chance of occurrence of each item being selected. Probability sampling increases your chances that the sample will be representative of the population.

SIMPLE RANDOM SAMPLING : The probability sampling process in which every individual or item from a population has the same chance of selection as every other individual or item. Every possible sample of a certain size has the same chance of being selected as every other sample that has that size.

SAMPLE SELECTION METHODS :

SAMPLING WITH REPLACEMENT : A sampling method in which each selected item is returned to the frame from which it was selected so that it has the same probability of being selected again.

SAMPLING WITHOUT REPLCAEMENT : A sampling method in which each selected item is not returned to the frame from which it was selected. Using this technique, an item can be selected no more than one time.

THE BAR GRAPH:

A chart containing rectangles (“bars”) in which the length of each bar represents the count, amount, or percentage of responses of one category.

PIE CHART:

A circle chart in which wedge-shaped areas—pie slices—represent the count, amount, or percentage of each category and the entire circle (“pie”) represents the total.

THE PARETO DIAGRAM:

A special type of bar chart in which the counts, amounts, or percentages of each category are presented in descending order left to right, along with a superimposed plotted line that represents a running cumulative percentage.

TWO WAY CROSS CLASSIFICATION TABLE :

A multicolumn table that presents the count or percentage of responses to two categorical variables. In two-way tables, the categories of one of the variables form the rows of the table, while the categories of the second variable form the columns. Cross-classification tables are also known as cross-tabulation tables.

HISTOGRAM : A special bar chart for grouped numerical data in which the frequencies or percentages of each group of numerical data are represented as individual bars on the vertical Y-axis and the variable is plotted on the horizontal X-axis. In a histogram, there are no gaps between adjacent bars as there would be in a bar chart of categorical data.

THE DOT SCALE DIAGRAM :

A chart in which each response is represented as a dot above a horizontal line that extends through the range of all values. Should two or more response values be identical, the dots for these responses are stacked (placed vertically) above each other.

THE TIME-SERIES PLOT:

A chart in which each point represents the value of a numerical variable at a specific time. By convention, the X-axis (the horizontal axis) always represents units of time, and the Y-axis (the vertical axis) always represents units of the variable.

THE SCATTER PLOT:

A chart that plots the values of two numerical variables for each response. In a scatter plot, the X-axis (the horizontal axis) always represents units of one variable, and the Y-axis (the vertical axis) always represents units of the second variable.

**Descriptive Statistics For Numerical Variables**

The Mean :

A number equal to the sum of the data values for a variable, divided by the number of data values that were summed.

Many sports statistics (including baseball batting averages and football yards per reception), average SAT score for incoming freshmen at a college, average age of the workers in a company, average waiting times at a bank.

The mean represents a “balance point” in a set of data values, similar to a fulcrum on a seesaw. As the only measure of central tendency that uses all the data values in a sample or population, the mean has one great weakness: individual extreme values can distort the balance point.

The Median :

The middle value in a set of data values for a variable when the data values have been ordered from lowest to highest value. When the number of data values to be summarized is even, you perform a special calculation to determine the median because data sets with an even number of values have no natural middle value