Statistics and Data exploration

Agenda

- **Why Study Statistics**
- **Application of Statistical Concepts in the Business World**
- **Types of Statistics**
- **Terminologies**
- **Population vs. Sample**
- **Types of Data**
- **Types of Statistical Variables**
- **Summarize the Data**
- **Understanding Summary Statistics**

Why Study Statistics?

- Data are everywhere
- Statistical techniques are useful in making many crucial decisions that impact our lives
- Irrespective of your career, you will make professional decisions that involve data
- Knowledge of statistical methods will contribute in making these decisions

Applications of Statistical Concepts in the Business World

- Finance – correlation and regression, index numbers, time series analysis
- Marketing – hypothesis testing, chi-square tests, nonparametric statistics
- Personnel – hypothesis testing, chi-square tests, nonparametric tests
- Operating management – hypothesis testing, estimation, analysis of variance,

Statistics

- The science of collecting, organizing, presenting, analyzing, and interpreting
- data to assist in making more elective decisions
- Statistical analysis is used to manipulate summarize, and investigate data, so as to get useful decision information results
- Statistics is the fun and is concerned with finding patterns in data

Types of Statistics

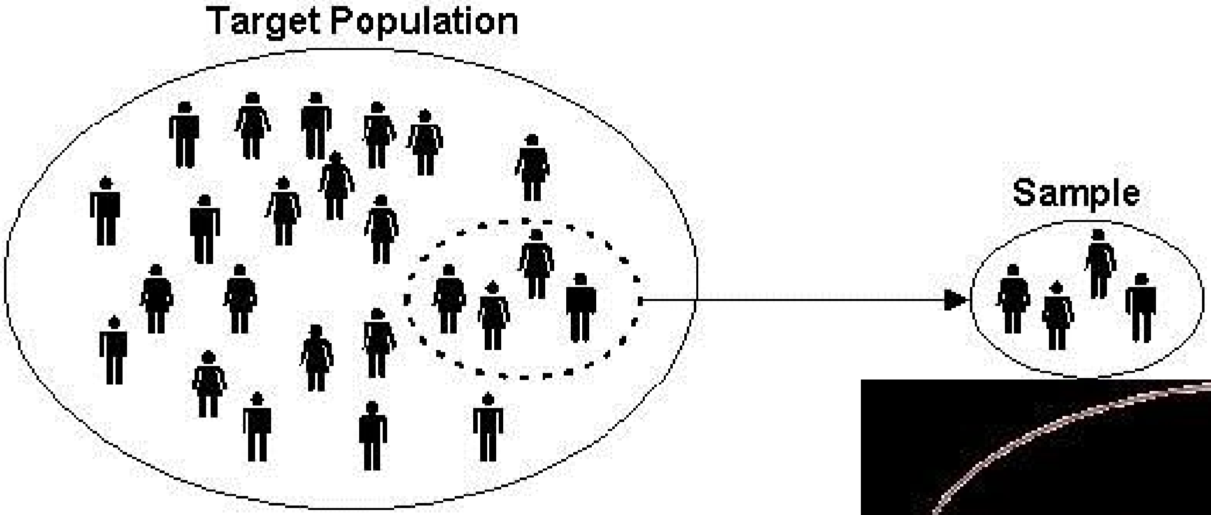
- Descriptive statistics – Methods of organizing, summarizing, and presenting data in an informative way
- Inferential statistics – tries to infer information about a population by using Information gathered by sampling.

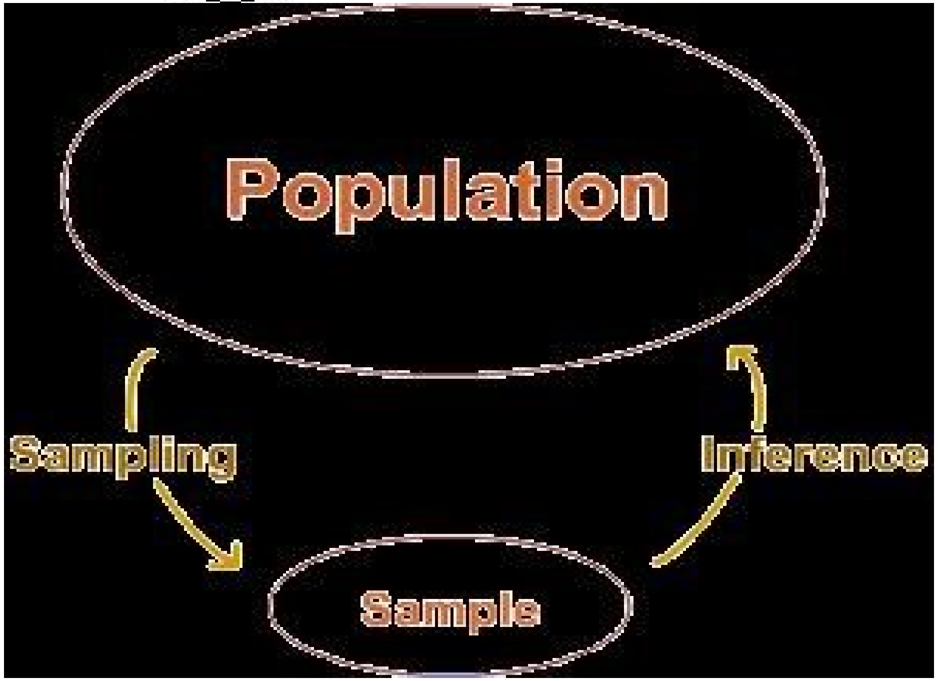
Terminology

- Population –The complete set of data elements is termed the population
- Sample – A proton, or part, of the population of interest selected for further analysis
- Parameter: A parameter is a characteristic of the whole population
- Statistic: A statistic is a characteristic of a sample, presumably measurable

Parameter is to Population as Statistic is to Sample

Terminology





Descriptive Statistics

- Descriptive statistics are numbers that are used to summarize and describe data
- The average or measure of center, consisting of the mean, median, mode or spread of the data
- Data - refers to the information that has been collected from an experiment, a survey, a historical record etc.
- E.g. of descriptive statistics- Number of people from India who won gold medal

Inferential Statistics

- Inferential statistics is concerned with making predictions or inferences about a population from observations and analyses of a sample.
- Two sections of Inferential Statistics –
  - Confidence Interval
  - Test of Significance (Hypothesis Testing)

Sampling

- A sample should be representative of the population i.e. it should have the same characteristics as the population it is representing.
- Sampling can be:
  - with replacement: a member of the population may be chosen more than once (picking a fruit from the fruit basket)
  - without replacement: a member of the population may be chosen only once (movie ticket)

Sampling Methods

- Sampling methods can be:
  - Probability Samples -random selection. More specifically, each sample from the population of interest has a known probability of selection under a given sampling scheme
  - Non Probability Samples - non random selection. You do not know the likelihood that any element of a population will be selected for study

Probability Samples Methods

Simple random sample	each sample of the same size has an equal chance of being selected.
Stratified sample	divide the population into groups called strata and then take a sample from each stratum.
Cluster sample	divide the population into strata and then randomly select some of the strata. All the members from these strata are in the cluster sample.
Systematic sample	randomly select a starting point and take every n-th piece of data from a listing of the population.

Statistical Data

- Data collection is commonly the most difficult, expensive, and time-consuming part of the entre research project
- Types of Data collected
  - Primary data are collected specifically for the analysis desired
  - Secondary data have already been compiled and are available for analysis

Data

Most of the data can be put into the following categories:

- Qualitative - data are measurements that each fall into one of several categories (skin color, religion, demographic data and other attributes of the population)
- Quantitative - data are observations that are measured on a numerical scale (distance traveled from village to city, number of children in the party, etc.)

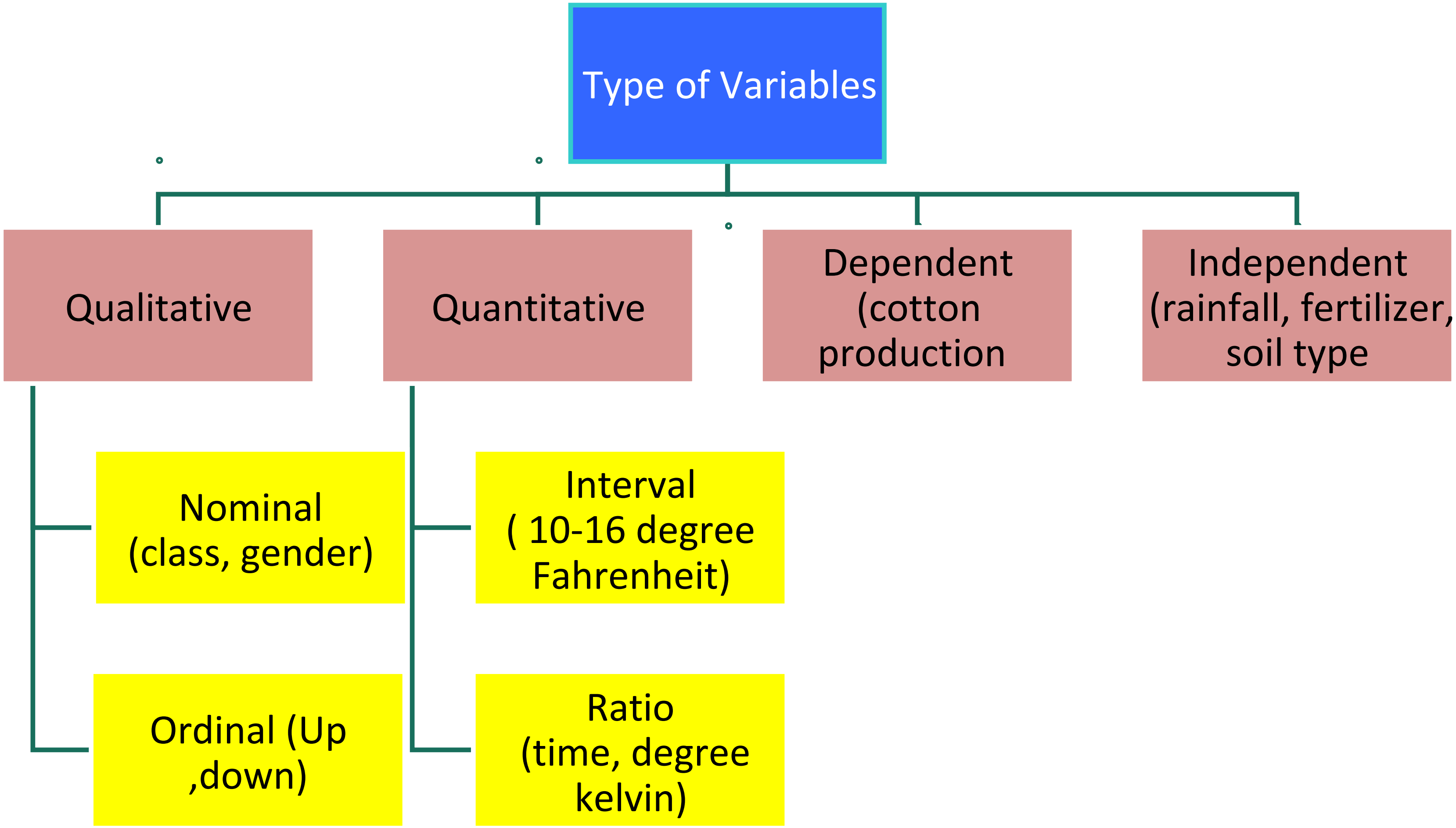
Qualitative Data

- Qualitative data is generally described by words or letters. They are not as widely used as quantitative data because many numerical techniques do not apply to the qualitative data. For example, it does not make sense to find an average hair color or blood type.
- Qualitative data can be separated into two subgroups:
  - dichotomic (if it takes the form of a word with two options (gender - male or female, binary – yes or no)
  - polynomic (if it takes the form of a word with more than two options (education – high school, under grad and post grad).

Quantitative Data

- Quantitative data is always numbers and is the result of counting or measuring attributes of a population.
- Quantitative data can be separated into two subgroups:
  - discrete: if it is the result of counting (the number of students of a given ethnic group in a class, the number of books on a shelf, etc.)
  - Continuous: if it is the result of measuring (distance traveled, weight of luggage, etc.)

Types of Statistical Variables



Type of Variables

Qualitative

Quantitative

Dependent (cotton production

Independent (rainfall, fertilizer, soil type

Nominal (class, gender)

Ordinal (Up ,down)

Interval (10-16 degree Fahrenheit)

Ratio (time, degree kelvin)

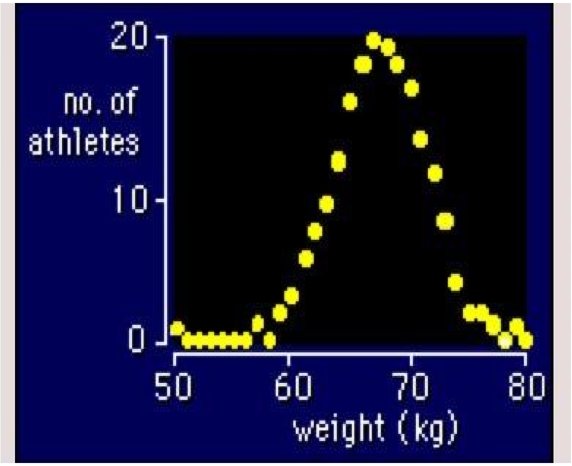
Scale of Measurement Explained

Nominal	Consist of categories in each of which the number of respective observations are recorded. The categories are in no logical order and have no particular relationship. The categories are said to be mutually exclusive since an individual, object or measurement can be included in only one of them.
Ordinal	Contain more information. Consists of distinct categories in which order is implied. Values in one category are larger or smaller than values in other categories (e.g. rating- excellent, good, fair, poor).
Interval	Is a set of numerical measurements in which the distance between numbers is of a known, constant size.
Ratio	Consists of numerical measurements where the distance between numbers is of a known, constant size. In addition, there is a no arbitrary zero point.

Summarize the data

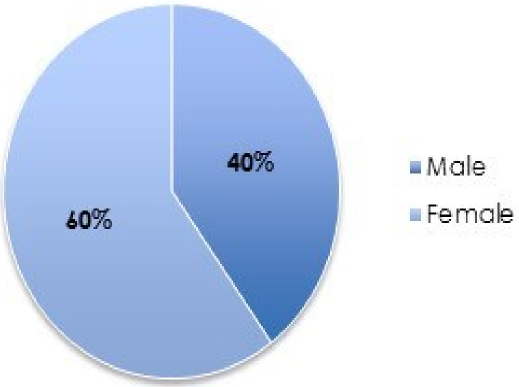
Summarize the Data

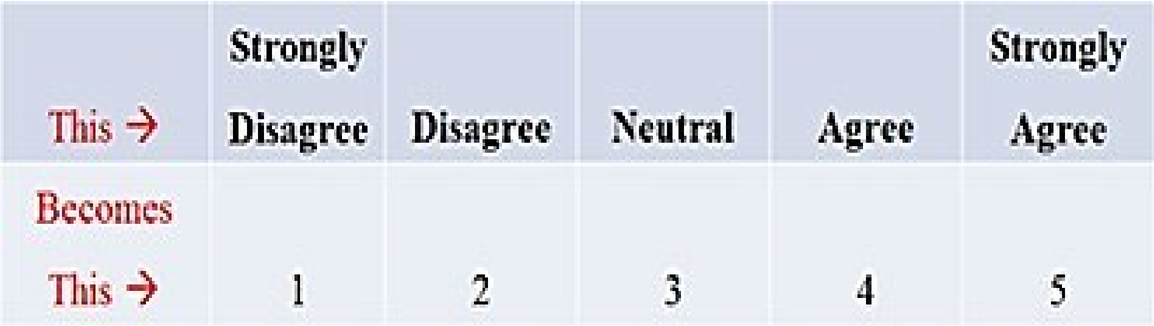
- People hate numbers, and they can't understand them in bulk. That's why you have to summarize data when you present results of your research
- Frequency Distributions - For numeric variables, one important way to summarize the values is to graph them as a frequency distribution. Here's what the weights of 200 athletes might look like in a frequency distribution done as a plot, which shows a point for the number of times each weight occurs
- You can also show the frequencies as vertical bars rather than points called as histogram



Summarize the Data

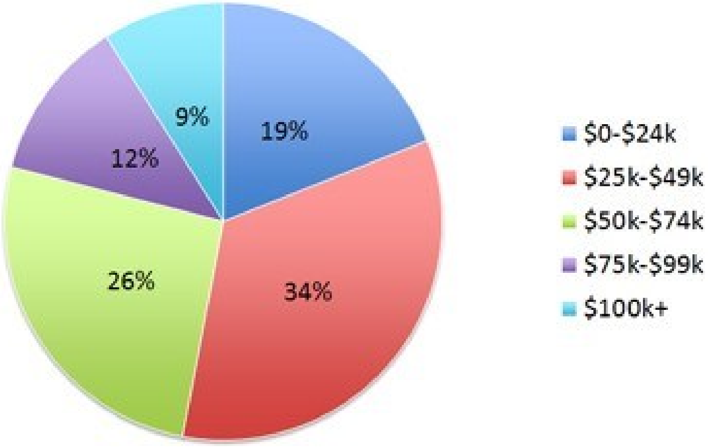
- Binary Responses
- If a question has only two possible response options
- (e.g., Male/Female, Yes/No, Agree/Disagree) such as
- the percent of women who responded, you can use
- the ubiquitous pie graph.
- Rating Scales
- Rating scale questions can be those that explicitly ask participants to rate their

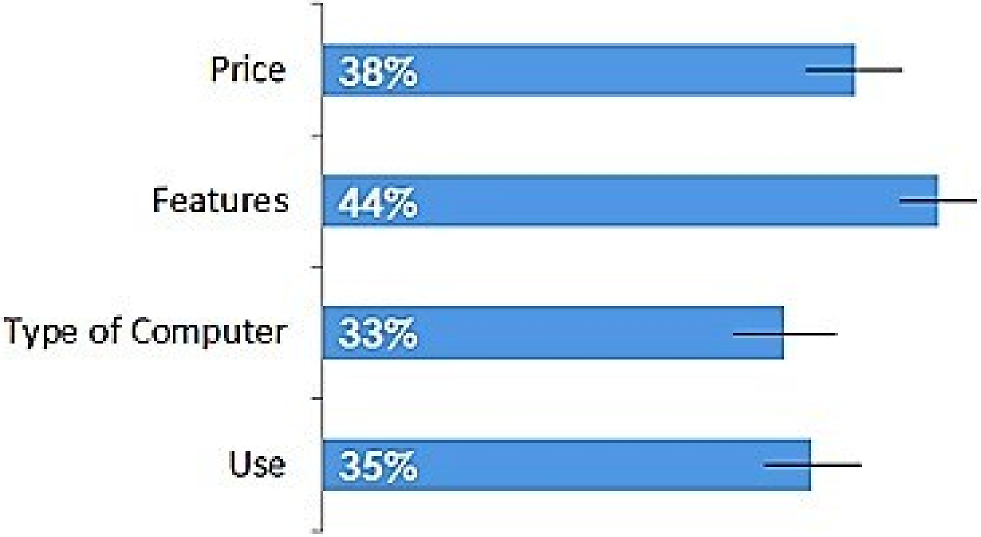


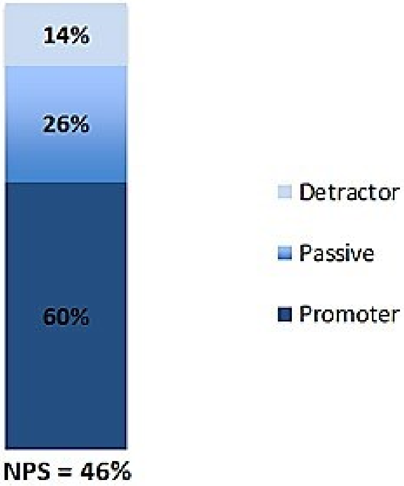


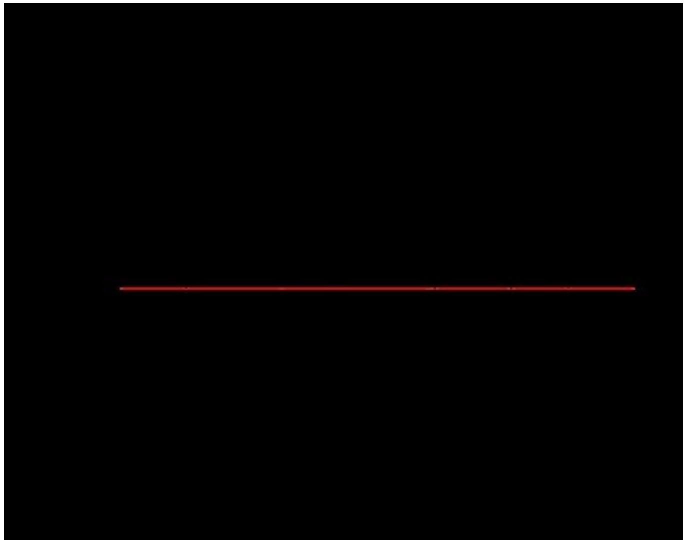
This →	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Becomes This →	1	2	3	4	5

Other Representations to Summarize Data









R code for Summary Charts

Frequency table

- – table (variable)
- Bar chart
- – barplot(table(variable))
- Histogram
- – hist(variable)
- Box-plot
- – boxplot(variable,horizontal=TRUE)

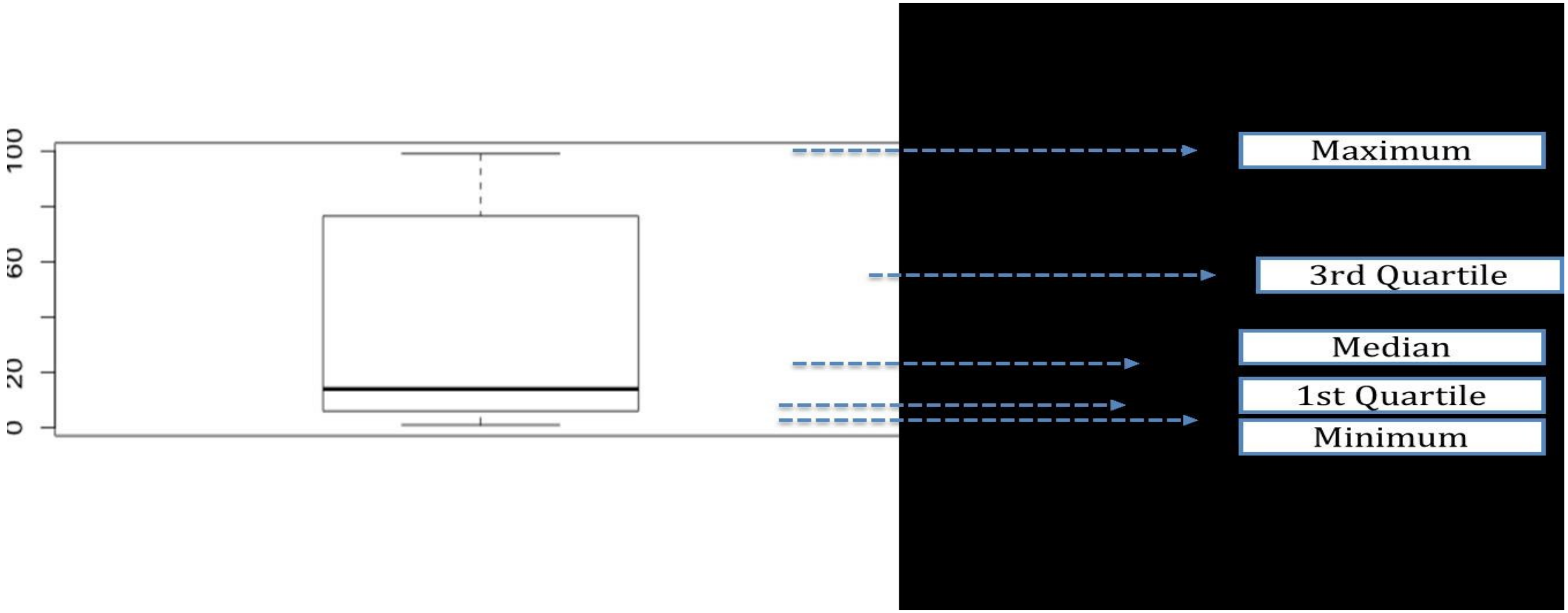
Understanding summary statistics

Understanding Summary Statistics

Mean	The arithmetic mean, is simply the arithmetic average of a group of numbers in the data set.
Mode	The mode is the most common or "most frequent" value in a data set.
Median	The median is the "middle value" in a set.
Quartile	It is a useful concept in statistics and is conceptually similar to the median. The first quartile is the data point at the 25th percentile, and the third quartile is the data point at the 75th percentile. The 50th percentile is the median.

Understanding Summary Statistics

```
d <- c(1,2,10,14,68,85,99)
> d
[1]  1  2 10 14 68 85 99
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.00    6.00   14.00   39.86   76.50   99.00 
> boxplot(d)
```



Maximum

3rd Quartile

Median

1st Quartile

Minimum

Understanding Summary Statistics

What quartiles do?

- The first quartile means that 25% of the data is smaller than the first quartile and 75% of the data is larger than this
- Similarly, in case of the third quartile, 25% of the data is larger than this, while 75% of it is smaller
- For the second quartile, which is nothing but the median, 50% or half of the data is smaller while half of the data is larger than this value
- Quartiles help us measure how the data is distributed in the two arms on either side of the median
- Thus if the first quartile is far away from the median while the third quartile is closer to it, it means that the data points that are smaller than the median are spread far apart while the data points that are greater than the median are closely packed together