Statistics

Chapter 1

The Nature of Probability and Statistics

Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.

Main Areas of Statistics

Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

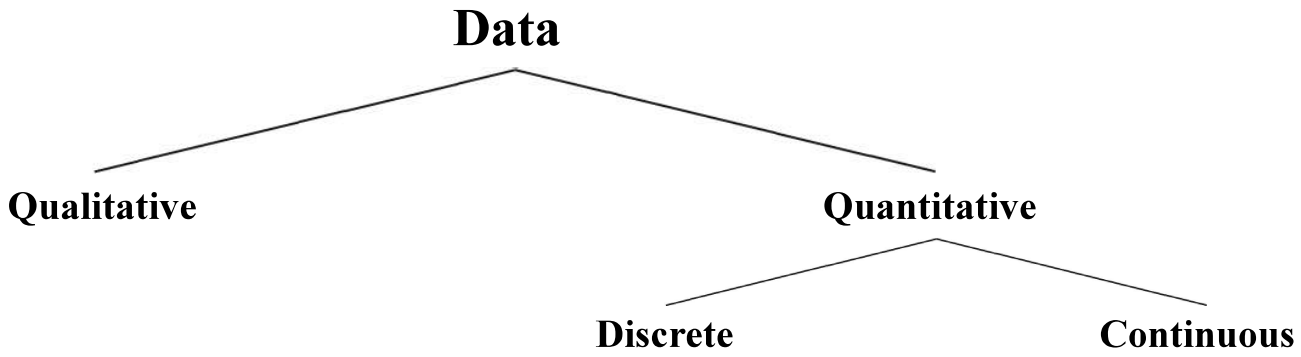
Inferential statistics consists of *generalizing* from samples to populations, performing estimation and hypothesis tests, determining relationships among variables, and making predictions.

Inferential statistics uses **probability** (the likelihood of an outcome occurring) to make conclusions and predictions.

A **population** consists of all subjects that are being studied.

A **sample** is a group of subjects selected from a population.

Data are the values (measurements or observations) that the variables can assume.



A **data set** is a collection of **data values**.

A **variable** is a characteristic or attribute that can assume different values.

A **random variable** is a variable whose values are determined by chance.

Types of Variables

Qualitative variables are variables that can be placed into categories, according to some characteristic or attribute.

Quantitative variables are numeric in nature and can be ordered or ranked. Quantitative variables can be either **discrete** or **continuous**.

A **continuous variable** is a quantitative variable that can assume ANY numerical value between any two specific values.

- Obtained by measuring.
- May include fractions and decimals.

A discrete variable is a quantitative variable that has either a finite number of possible values or a countable number of possible values.

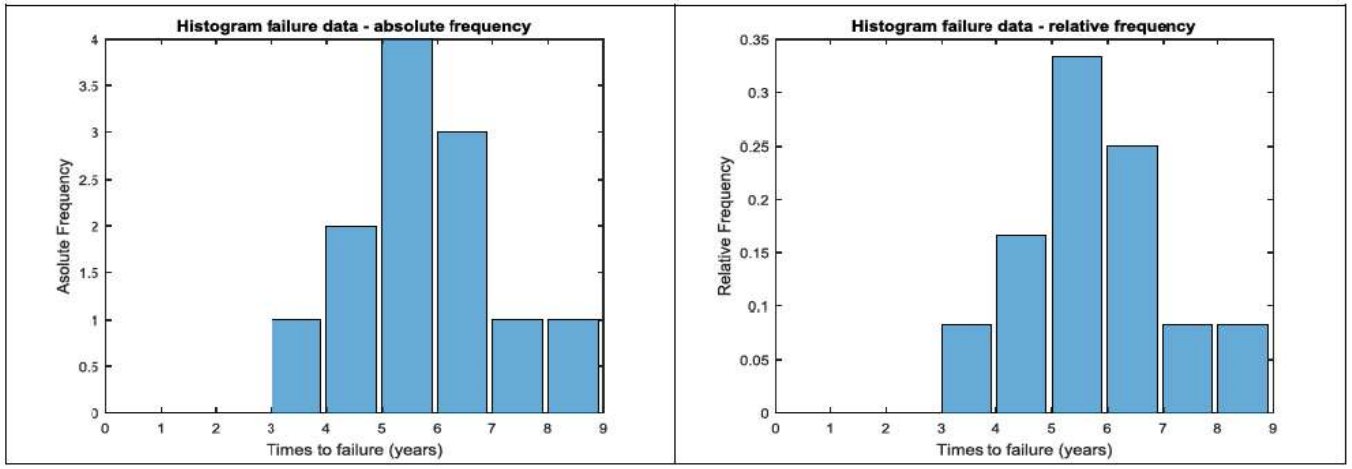
- Countable means that the values result from counting, such as 0, 1, 2, 3…

Histogram

Histogram: set of columns describing the distribution of specific data. Each column represents a number of items or frequency.
This graphic gives an idea of sample shape.

Absolute frequency: It indicates the number of data points.

Relative frequency: I t indicates the percentage of the total number of data points.



The characteristics of any data can be summarized by knowing some statistical properties:

1. The data center is measured using the median or mean.
2. Data spread is measured using variance and standard deviation.
3. Knowing that most of the data is located using the quartiles.
4. Find outlie

Median

When arranging the data in ascending or descending order, the median is the statement in which 50% of the data fall before and 50% after it, and it is calculated:

$$X(1), X(2), ..., X(n-1), X(n)$$

$$\tilde{X} = \begin{cases} X(\frac{n+1}{2}) & \text{if } n \text{ is odd} \\ \frac{X(\frac{n}{2}) + X(\frac{n}{2}+1)}{2} & \text{if } n \text{ is even} \end{cases}$$

Example 1: Find the median for the data listed below 60 72 40 80 63.

Sol:

We arrange the data in ascending order

40 60 63 72 80

Since the number of data is odd, the median value (third) represents the median

$\tilde{X} = X\ (5 + 1)/2 = X3 = 63$

Example 2: Find the median for the data listed below 72 60 72 40 80 63.

Sol:

We arrange the data in ascending order

40 60 63 72 72 80

Since the number of data is even, the median value is the average of the two readings in the middle of the data

$\tilde{X} = (63+72)/2 = 67.5$

Statistics

Chapter 1

The Nature of Probability and Statistics

Data Collection and Sampling Techniques

Data is often collected via surveys.

1. Telephone Surveys
2. Mailed Questionnaire
3. Personal Interview
4. Internet Survey

- What are advantages and disadvantages of data collection through surveys?

	Advantages	Disadvantages
Telephone Surveys	1. Less costly than personal interview. 2. People may be more candid.	1.Not everyone has a phone. 2. Cell phones typically not included. 3.Tone of interviewer's voice may affect response.
Mailed Questionnaire	Can cover wider geographic area than phone survey or personal interview 2. Respondents can remain anonymous 3. Less expensive	1.Low number of responses 2.Inappropriate answers to questions 3.Low reading abilities or not understanding questions may create useless responses.

	Advantages	Disadvantages
Personal Interview	In-depth responses to questions	1.Interviewers must be trained in asking questions and recording responses (which is costly). 2. Interviewer may be biased in the selection of participants.
Internet survey	As with telephone and mail surveys, 1. Inexpensive, often free 2. Candor 3. Large geographic coverage 4. Anonymity	1.Will miss demographics without computer access 2. May have inappropriate answers if questions are misunderstood

Sampling Techniques

- Researchers use samples to collect data and information about a particular variable from a population.
- Samples save time, money, and may actually allow a researcher to collect better information.
- Samples need to be representative of the population or they are meaningless in drawing conclusions about the population.
- Sampling must be done in a way that the samples are **unbiased**—that each subject in the population has an equal chance of being in the sample.

Technique	Description
Random sampling	Uses chance methods or random numbers to select the sample. **Everyone or everything from the population has the same chance of being selected for the sample** and it is the best way of obtaining a representative sample.
Systematic sampling	Numbers each subject of the population and then **selects every kth subject.**
Convenience sampling	Selects subjects that are convenient for the researcher. These samples are typically of not statistical value.
Stratified sampling	Divides the population into groups (called strata) according to some characteristic that is important to the study, then randomly samples subject from each group.
Cluster Sampling	Divides the population into groups called clusters by some means such as geographic area or schools in a school district, etc. Then randomly select some of the clusters and use ALL members of the selected clusters.

Observational and Experimental Studies

Observational study - the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.

Experimental study - the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.