STA250 Probability and Statistics

Lecture 1: Introduction to Statistics and Data Analysis

Asst. Prof. Abdullah YALÇINKAYA

Ankara University, Faculty of Science, Department of Statistics

ayalcinkaya@ankara.edu.tr

_2023_

How to Contact?

- https://ekampus.ankara.edu.tr/
- Microsoft Teams (Online Class / Chat)
- E-mail: ayalcinkaya@ankara.edu.tr

Breakdown of Grades

Here is the plan:

The grade will be composed of the grades on:

- One midterm (30%),
- Final exam (80%).

STA250 Probability and Statistics

Reference Books

This lecture notes are prepared according to the contents of

«PROBABILITY & STATISTICS FOR ENGINEERS & SCIENTISTS»
by Walpole, Myers, Myers and Ye

«APPLIED STATISTICS AND PROBABILITY FOR ENGINEERS»
by Montgomery and Runger

«Statistics for Biomedical Engineers and Scientists How to Visualize and Analyze Data» by Andrew P. King and Robert J. Eckersley

Course Content

Week 1. Introduction To Statistics And Data Analysis

Week 2. Summarizing Data: Tables, Diagrams And Graphs

Week 3. Summarizing Data: Measures Of Tendency And Dispersion

Week 4. Probability

Week 5. Discrete Random Variables And Their Probability Distributions

Week 6. Continuous Random Variables And Their Probability Distributions

Week 7. Sampling Distributions and Central Limit Theorem

Week 8. Properties of Point Estimators and Methods of Estimation.

Week 9-10. Hypothesis Testing

Week 11-12. Simple Linear Regression and Correlation

The role of Statistics in Engineering

The Engineering Method and Statistical Thinking

- Engineers solve problems of interest to society by the efficient application of scientific principles.
- The engineering or scientific method is the approach to formulating and solving these problems.

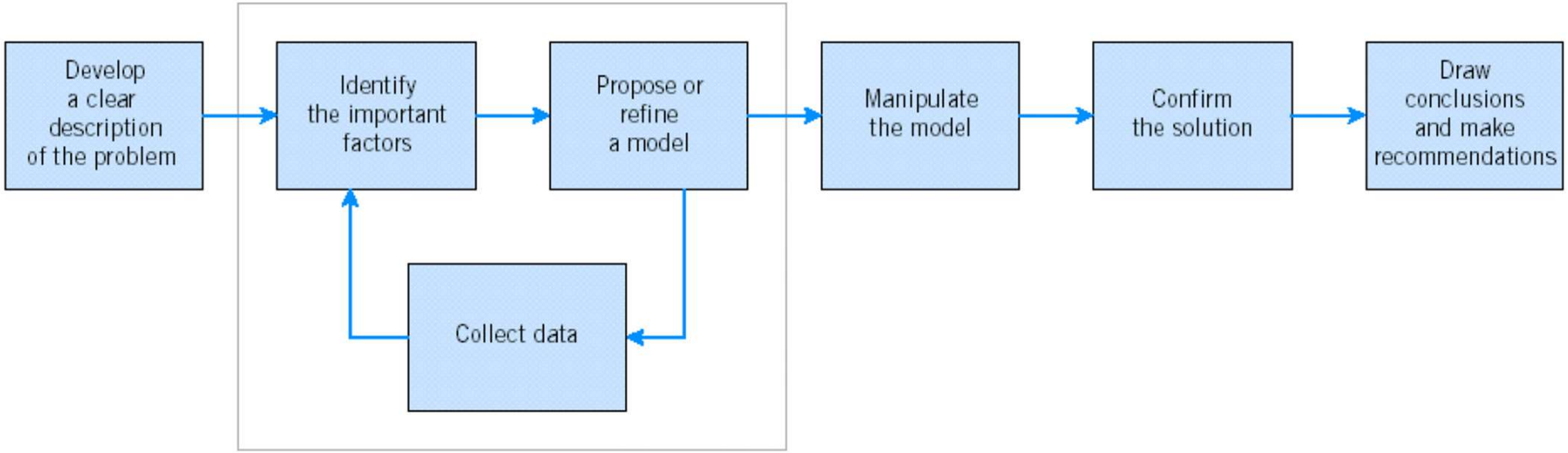

Develop a clear description of the problem
Identify the important factors
Propose or refine a model
Manipulate the model
Confirm the solution
Draw conclusions and make recommendations
Collect data

Figure 1-1 The engineering problem-solving method.

The Engineering Method and Statistical Thinking

- **The field of Probability**
  - Used to quantify likelihood or chance
  - Used to represent risk or uncertainty in engineering applications
  - Can be interpreted as our degree of belief or relative frequency
- **The field of Statistics**
  - Deals with the collection, presentation, analysis, and use of data to
    - make decisions
    - solve problems.

Definitions

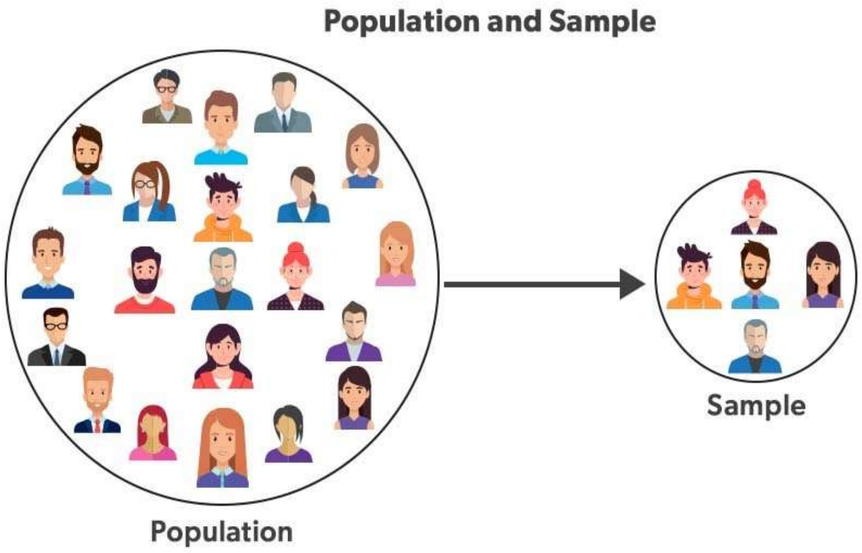
- **Statistics is the science of**
  - collection of methods for planning experiments,
  - obtaining data, and then organizing,
  - summarizing, presenting,
  - analyzing, interpreting,
  - drawing conclusions.

The Engineering Method and Statistical Thinking

- Statistical techniques are useful for describing and understanding variability.
- By variability, we mean successive observations of a system or phenomenon do not produce exactly the same result.
- Statistics gives us a framework for describing this variability and for learning about potential sources of variability.

Definitions

- **Population**
  - All subjects possessing a common characteristic that is being studied.
- **Sample**
  - A subgroup or subset of the population.



Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.

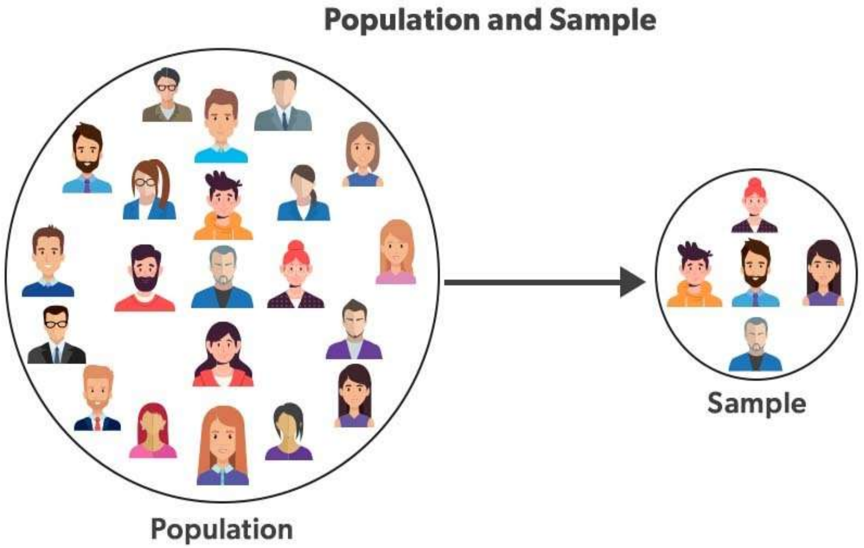
The term **sample size** simply means the number of elements in the sample.

Often in statistics, we compare samples from two different populations and try to determine statistically if the populations are significantly different.

Population and Sample Examples

All the students in the class are population whereas the top 10 students in the class are the sample.

All the members of the parliament is population and the female candidates present there is the sample.



Definitions: Sampling

- Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population.

- Obvious questions:

– How best to obtain the sample and make the observations?

– Once the sample data are in hand, how best to use them to estimate the characteristic of the whole population?

Definitions: Sampling

- Basically, there are two types of sampling.
- They are:
  - Probability sampling
  - Non-probability sampling

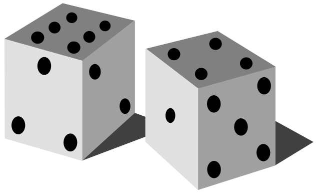
Probability Sampling

- In probability sampling, the population units **cannot be selected** at the discretion of the researcher. This can be dealt with following certain procedures which will ensure that **every unit of the population consists of one fixed probability** being included in the sample. Such a method is also called **random sampling**.

- Some of the techniques used for probability sampling are:
- Simple random sampling
- Systematic sampling
- Cluster sampling
- Stratified Sampling
- Disproportionate sampling
- Proportionate sampling
- Optimum allocation stratified sampling
- Multi-stage sampling

Definitions

- **Simple Random Sampling**
  - Every individual or item from the frame has an equal chance of being selected
  - Selection may be with replacement or without replacement
  - Samples obtained from table of random numbers or computer random number generators



Definitions

- **Systematic Sampling**
  - Sampling in which data is obtained by selecting every *k*th object.

Decide on sample size: **n**

Divide frame of **N** individuals into groups of **k** individuals: **k**=**N**/**n**

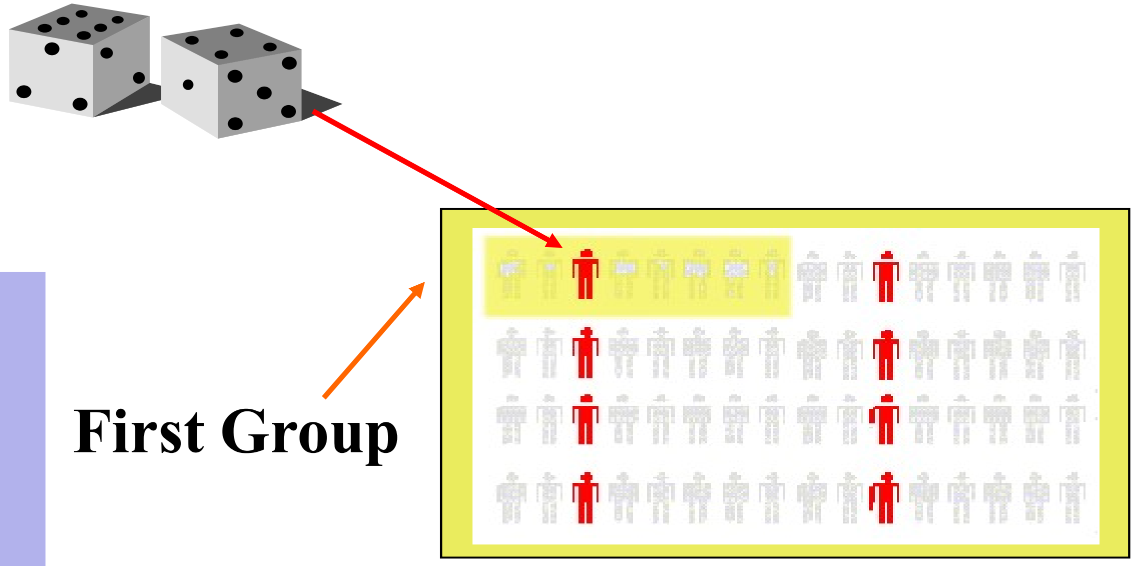
Randomly select one individual from the 1st group

Select every k-th individual thereafter

N = 64

n = 8

k = 8



First Group

Advantages-Disadvantages

- **Simple random sample and systematic sample**
  - Simple to use
  - May not be a good representation of the population’s underlying characteristics
- **Stratified sample**
  - Ensures representation of individuals across the entire population
- **Cluster sample**
  - More cost effective
  - Less efficient (need larger sample to acquire the same level of precision)

Non-Probability Sampling

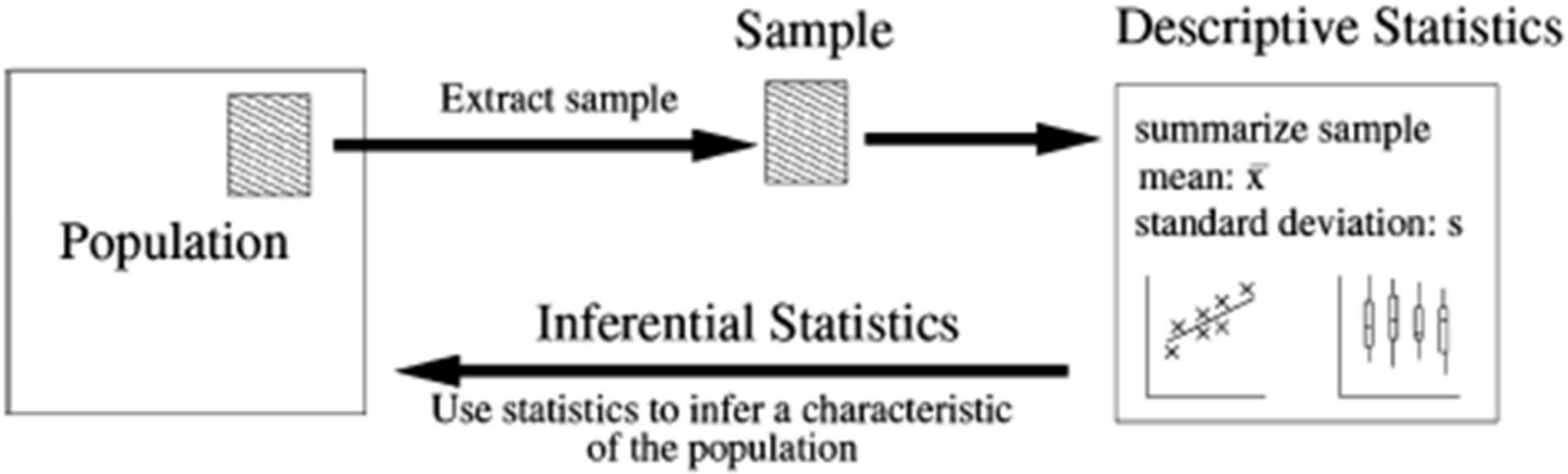
- In non-probability sampling, the population units **can be selected** at the discretion of the researcher.
- Those samples will use the **human judgements** for selecting units and **has no theoretical basis** for estimating the characteristics of the population.
- Some of the techniques used for non-probability sampling are
- Quota sampling
- Judgement sampling
- Purposive sampling

Definitions

- **Parameter**
  - Characteristic or measure obtained from a population.
- **Statistic (not to be confused with Statistics)**
  - Characteristic or measure obtained from a sample.

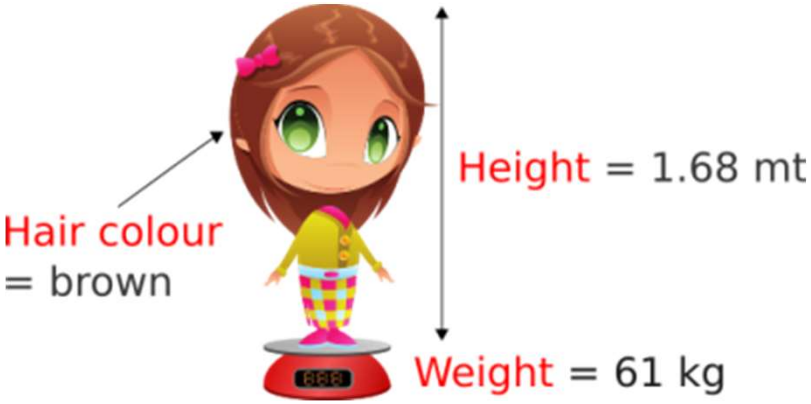
Definitions

- **Descriptive Statistics**
  - Collection, organization, summarization, and presentation of data.
- **Inferential Statistics**
  - Generalizing from samples to populations using probabilities. Performing hypothesis testing, determining relationships between variables, and making predictions.



Definitions

- **Variable**
  - Characteristic or attribute that can assume different values
- **Random Variable**
  - A variable whose values are determined by chance



Variables

- Concepts that are observable and measurable
- Have a dimension that can vary
- Narrow in meaning
- Examples:
  - Color classification
  - Loudness
  - Level of satisfaction/agreement
  - Amount of time spent
  - Media choice

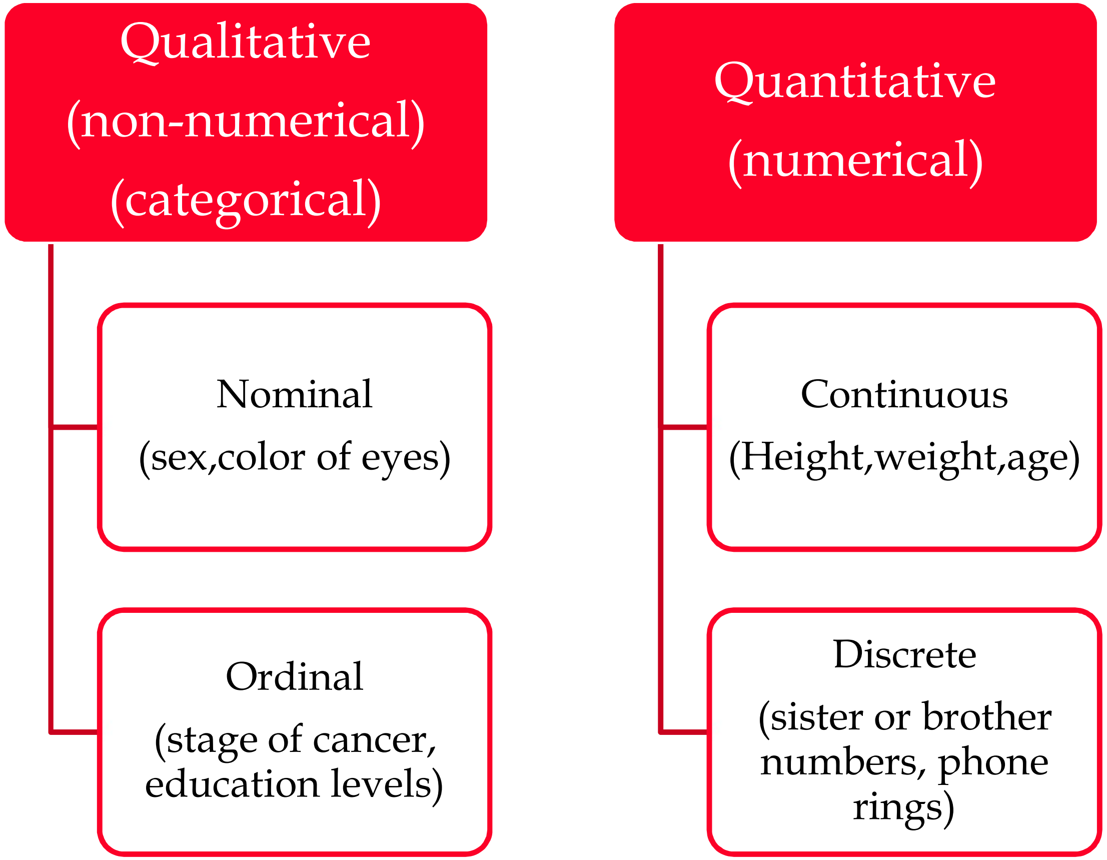
Definitions

A variable is any characteristic of an individual.

A variable can take different values for different individuals.

- **Qualitative Variables**
  - Variables which assume non-numerical values.
- **Quantitative Variable**
  - Variables which assume numerical values.
- **Discrete Variables**
  - Variables which assume a finite or countable number of possible values. Usually obtained by counting.
- **Continuous Variables**
  - Variables which assume an infinite number of possible values. Usually obtained by measurement.

Types of Variables


Qualitative (non-numerical) (categorical)

Nominal (sex,color of eyes)

Ordinal (stage of cancer, education levels)

Quantitative (numerical)

Continuous (Height,weight,age)

Discrete (sister or brother numbers, phone rings)

Numbers, numbers everywhere

555-867-5309

9001

9

.05

3.5

97.5

502

4,832

834,722

77

999

65.87

36^2

.998

4001

51

.56732

1,248,965

2,387

9

21

672

145

324

999-99-9999

409

35.5

Categorical variables

have values that describe a '**quality**' or '**characteristic**' of a data unit, like '**what type**' or '**which category**'.

Categorical variables may be further described as **ordinal** or **nominal**:

An **ordinal variable** is a categorical variable. Observations can take a value that can be logically ordered or ranked.

The categories associated with ordinal variables can be ranked higher or lower than another, but do not necessarily establish a numeric difference between each category.

Examples of ordinal categorical variables include **academic grades** (i.e. A, B, C), **clothing size** (i.e. small, medium, large, extra large) and **attitudes** (i.e. strongly agree, agree, disagree, strongly disagree).

A **nominal variable** is a categorical variable. Observations can take a value that is not able to be organised in a logical sequence.

Examples of nominal categorical variables include **sex**, **business type**, **eye colour**, **religion** and **brand**.

The data collected for a categorical variable are qualitative data.

Numeric variables

have values that describe a **measurable quantity** as a **number**,
like '**how many**' or '**how much**'.

Numeric variables may be further described as either **continuous** or **discrete**:

A **continuous variable** is a numeric variable. Observations can take any value between a certain set of real numbers.

The value given to an observation for a continuous variable can include values as small as the instrument of measurement allows. Examples of continuous variables include **height**, **time**, **age**, and **temperature**.

A **discrete variable** is a numeric variable. Observations can take a value based on a count from a set of distinct whole values.

A discrete variable cannot take the value of a fraction between one value and the next closest value.

Examples of discrete variables include the **number of registered cars**, **number of business locations**, and **number of children in a family**, all of which measured as whole units (i.e. 1, 2, 3 cars).

The data collected for a numeric variable are quantitative data.

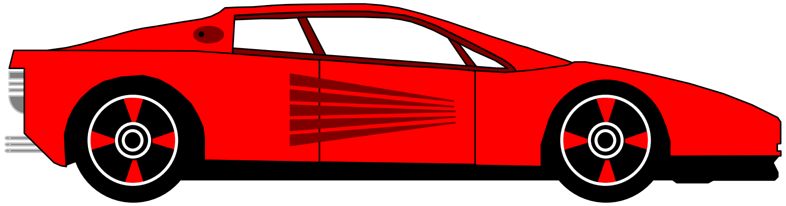
Scale

- It is the tools and equipment used to obtain numerical data
- Represents a composite measure of a variable
- Series of items arranged according to value for the purpose of quantification
- Provides a range of values that correspond to different characteristics or amounts of a characteristic exhibited in observing a concept.
- **Scales come in four different levels:**
- Nominal
- Ordinal
- Interval
- Ratio

Nominal Scale

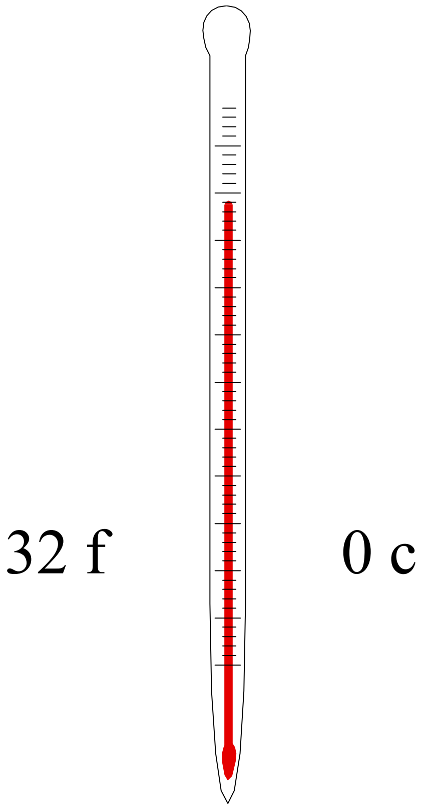
- Indicates a difference

Ordinal Scale



- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)

Interval Scale



- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)
- Indicates the amount of the difference (in equal intervals)

Ratio Scale

- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)
- Indicates the amount of the difference (in equal intervals)
- Indicates an absolute zero

- **The study of statistics has two major branches:**
  - descriptive (exploratory) statistics
  - inferential statistics.
- **Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data
- **Inferential statistics** is the branch of statistics that involves using a sample to draw conclusions about population. A basic tool in the study of inferential statistics is probability.

Example (Descriptive Statistics)

- □ **Collect data**
  - e.g. Survey
- □ **Present data**
  - e.g. Tables and graphs
- □ **Characterize data**
  - e.g. Sample mean = $\frac{\sum X_i}{n}$

Example (Inferential Statistics)

- **Estimation**
  - e.g.: Estimate the population mean weight using the sample mean weight
- **Hypothesis testing**
  - e.g.: Test the claim that the population mean weight is 120 pounds

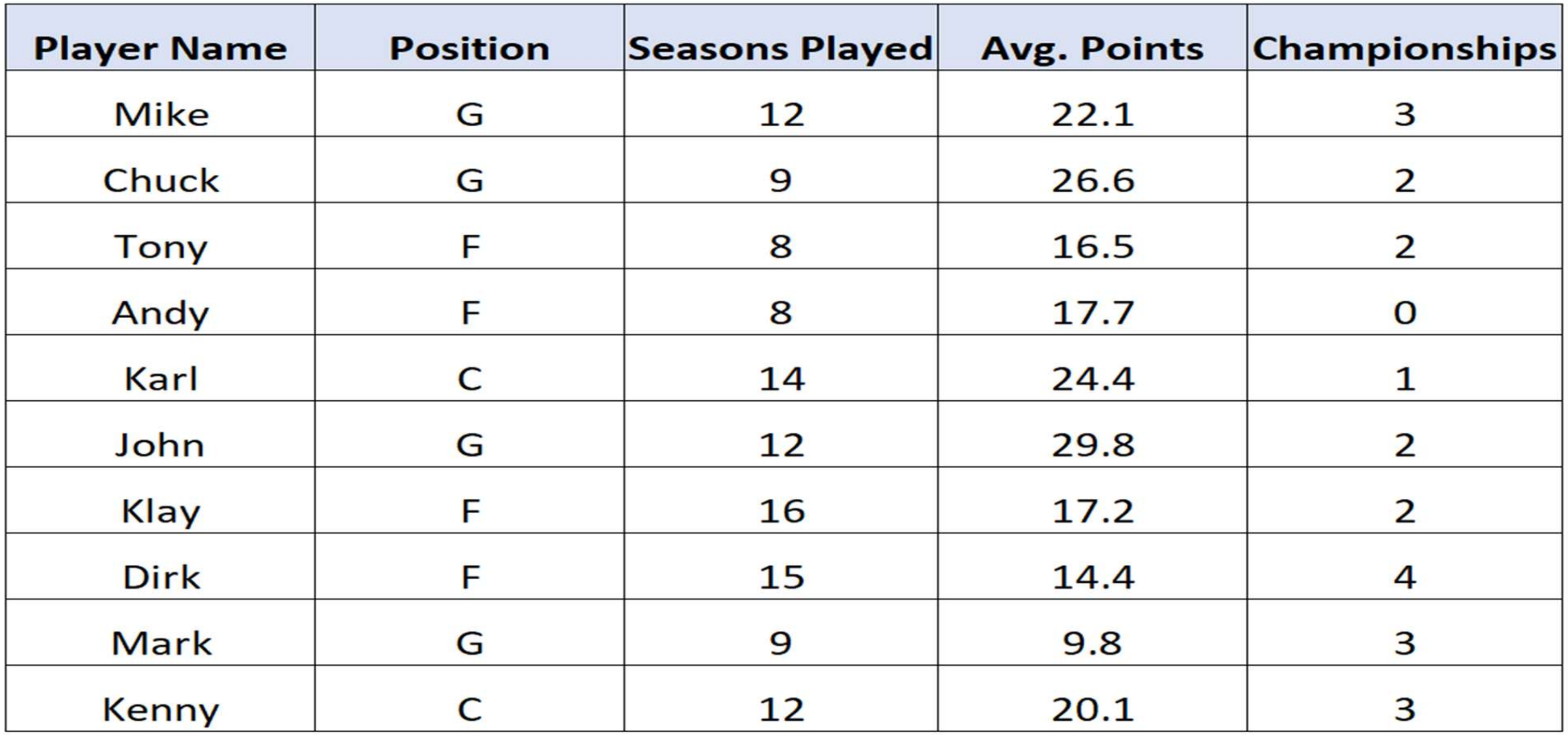
Drawing conclusions and/or making decisions concerning a population based on sample results.

Example (Type of Variables)

- **Consider the following dataset with information about 10 different basketball players:**

Variable Types:

Qualitative Qualitative Quantitative Quantitative Quantitative

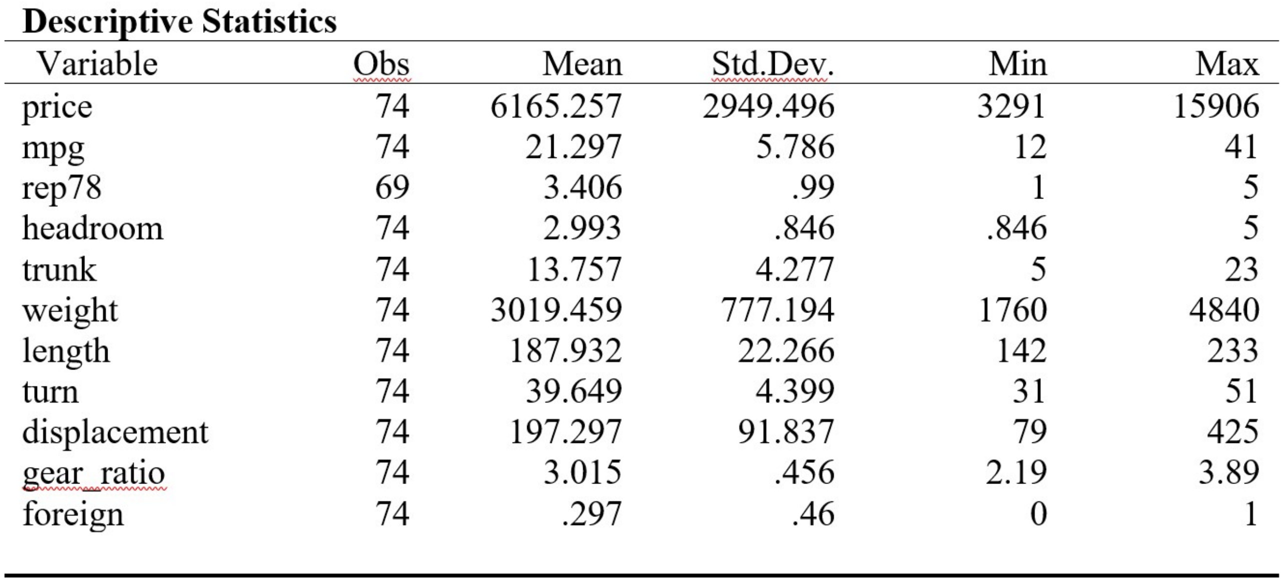


Player Name	Position	Seasons Played	Avg. Points	Championships
Mike	G	12	22.1	3
Chuck	G	9	26.6	2
Tony	F	8	16.5	2
Andy	F	8	17.7	0
Karl	C	14	24.4	1
John	G	12	29.8	2
Klay	F	16	17.2	2
Dirk	F	15	14.4	4
Mark	G	9	9.8	3
Kenny	C	12	20.1	3

Summarizing Data

□ Can be a table, graph or Numerical Measures

- Tables: A table is an arrangement of information in rows and columns containing cells that make comparing and contrasting information easier.



Descriptive Statistics

Variable	Obs	Mean	Std.Dev.	Min	Max
price	74	6165.257	2949.496	3291	15906
mpg	74	21.297	5.786	12	41
rep78	69	3.406	.99	1	5
headroom	74	2.993	.846	.846	5
trunk	74	13.757	4.277	5	23
weight	74	3019.459	777.194	1760	4840
length	74	187.932	22.266	142	233
turn	74	39.649	4.399	31	51
displacement	74	197.297	91.837	79	425
gear_ratio	74	3.015	.456	2.19	3.89
foreign	74	.297	.46	0	1

- Source:https://www.semanticscholar.org/

Summarizing Data

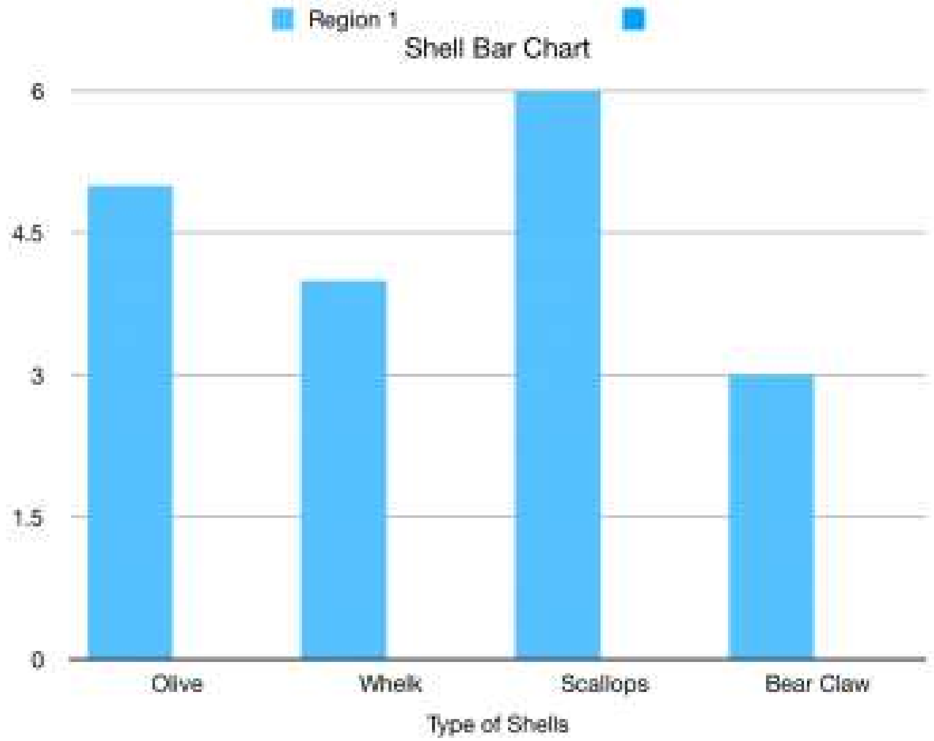
- Graph: It is a diagram showing the relationships between two or more variables.

- Categorical Variables
  - Pie
  - Bar

- Quantitativa Variables
  - Histogram
  - Stemplots (Stem-and-leaf plots)
  - Box Plot
  - Dot Diagram
  - Scatter Plot

Summarizing Data

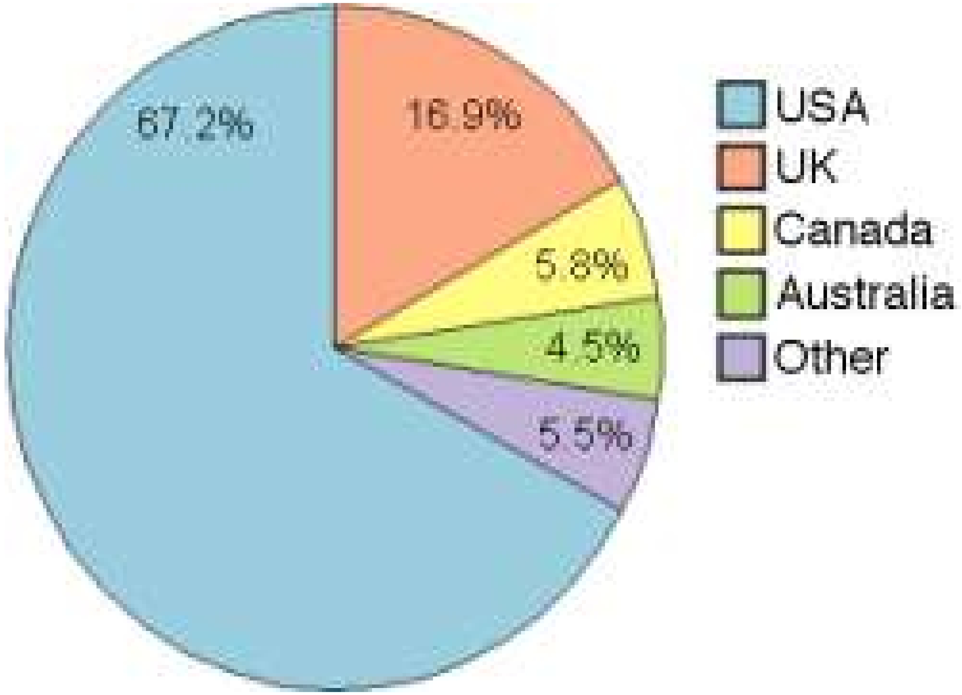
- **Bar Chart**



- Source : Wikipedia

Summarizing Data

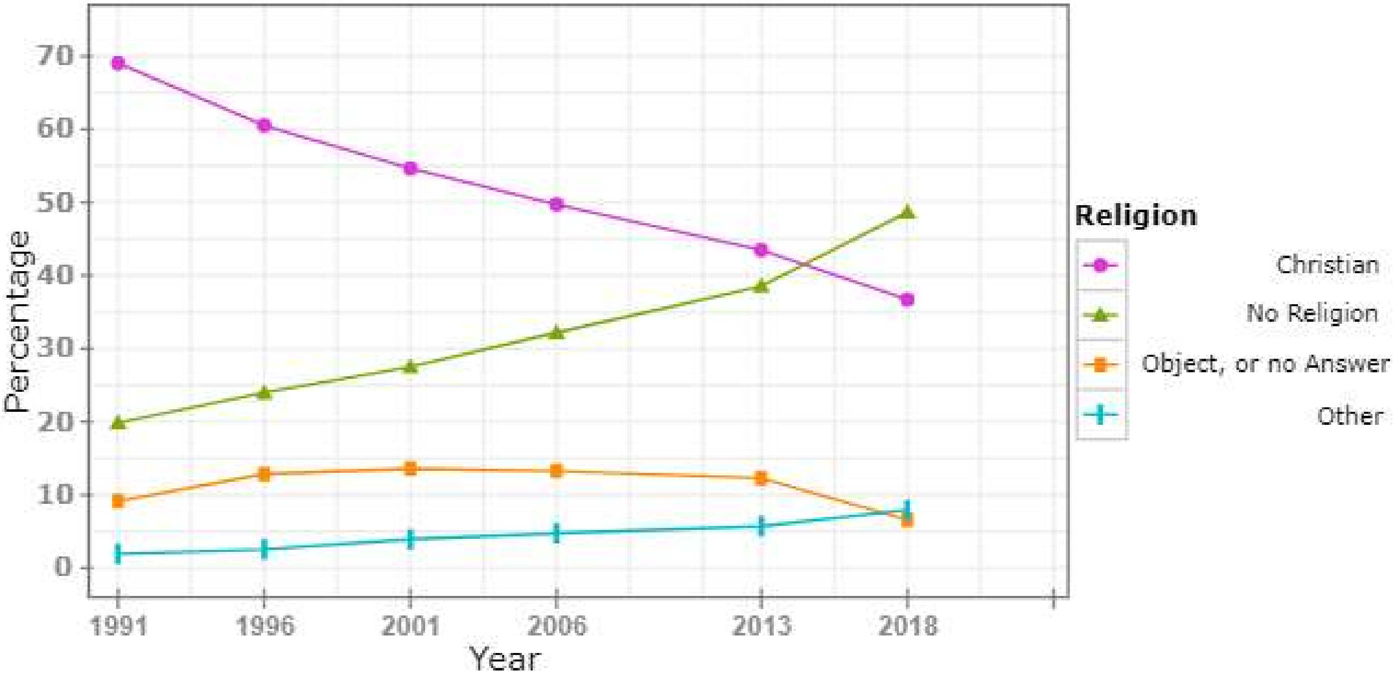
- **Pie Chart**



- Source : Wikipedia

Summarizing Data

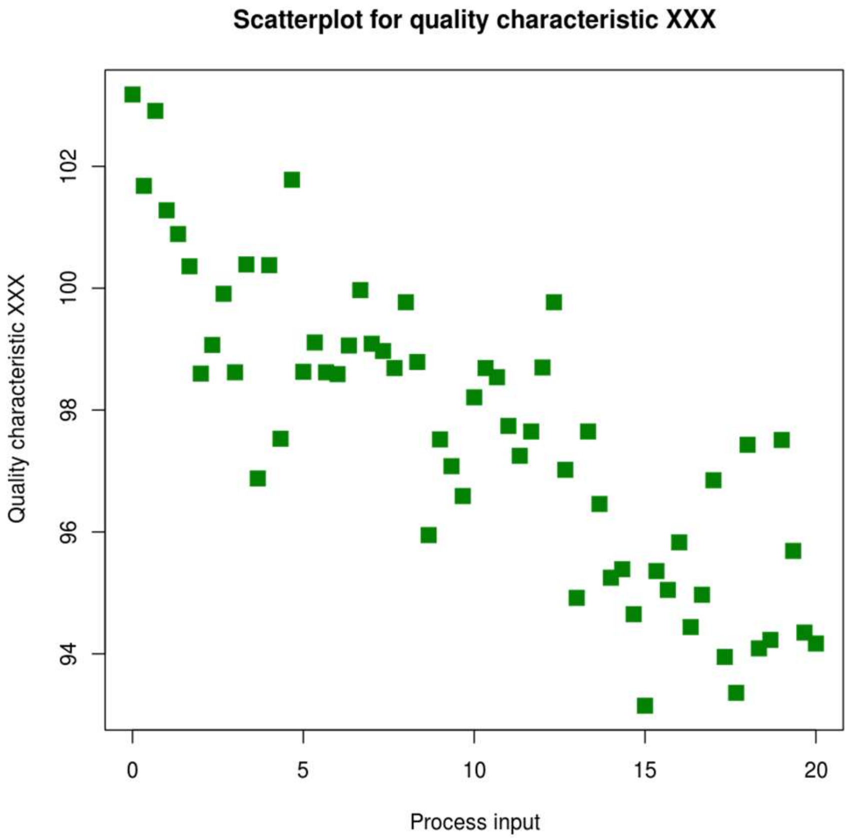
- **Line Graph**



- Source : Wikipedia

Summarizing Data

- **Scatter plot**



- Source : Wikipedia

Summarizing Data

Stem and Leaf Plot

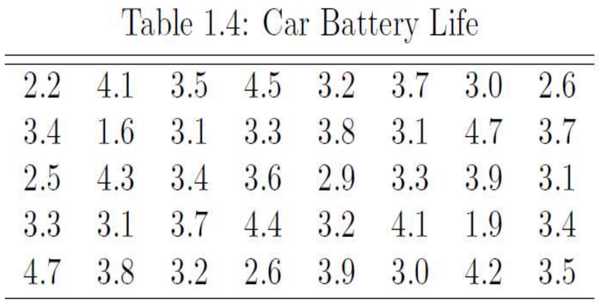


Table 1.4: Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

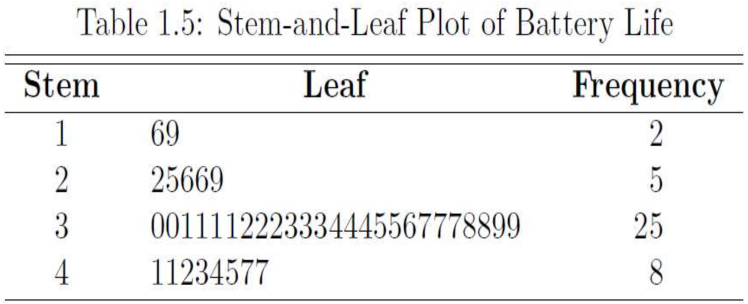


Table 1.5: Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1	69	2
2	25669	5
3	0011112223334445567778899	25
4	11234577	8

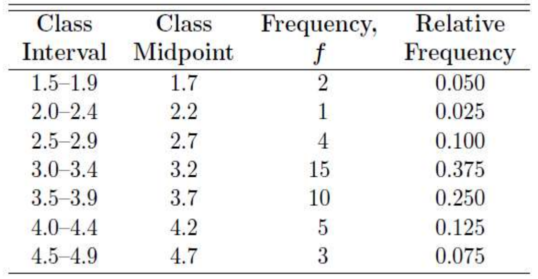
- A stem-and-leaf plot is a device for presenting quantitative data in a graphical format, similar to a histogram, to assist in visualizing the shape of a distribution.

- the Stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number.
- In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf.

Summarizing Data

Histogram

Table 1.7: Relative Frequency Distribution of Battery Life



Class Interval	Class Midpoint	Frequency, f	Relative Frequency
1.5–1.9	1.7	2	0.050
2.0–2.4	2.2	1	0.025
2.5–2.9	2.7	4	0.100
3.0–3.4	3.2	15	0.375
3.5–3.9	3.7	10	0.250
4.0–4.4	4.2	5	0.125
4.5–4.9	4.7	3	0.075

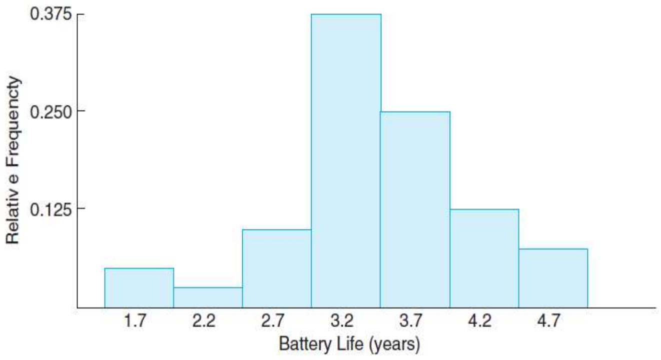


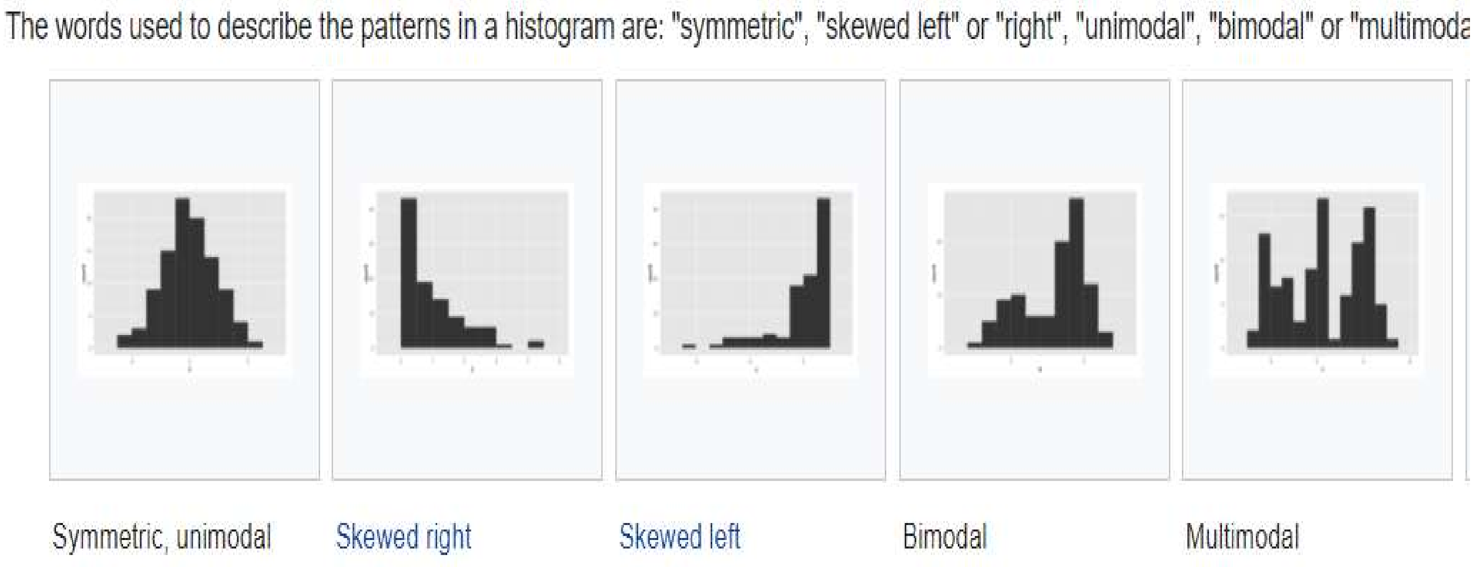
Figure 1.6: Relative frequency histogram.

- Given a sample of data points, we divide data into equally-spaced intervals, and count the number of data points that fall into each interval.
- A histogram is a bar chart with the length of each bar proportional to the number of observations in that interval.
- A histogram for a sample will be an approximation of the probability distribution of the population.

Summarizing Data

Shape of Distribution

- Histogram draw picture representing distribution.
- It may be symmetric, asymmetric or unimodal, bimodal



The words used to describe the patterns in a histogram are: "symmetric", "skewed left" or "right", "unimodal", "bimodal" or "multimoda

Symmetric, unimodal

Skewed right

Skewed left

Bimodal

Multimodal

Source: wikipedia

Summarizing Data

Shape of Distribution

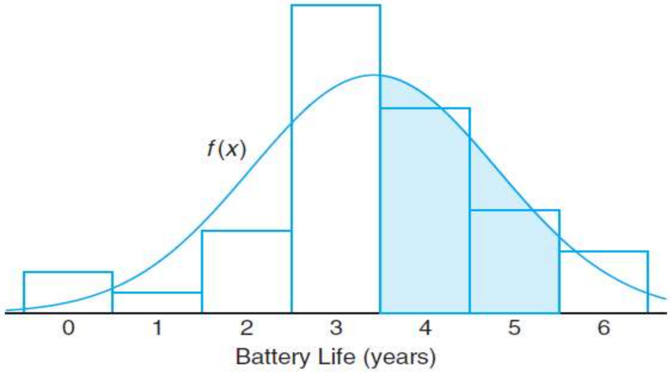


Figure 1.7: Estimating frequency distribution.

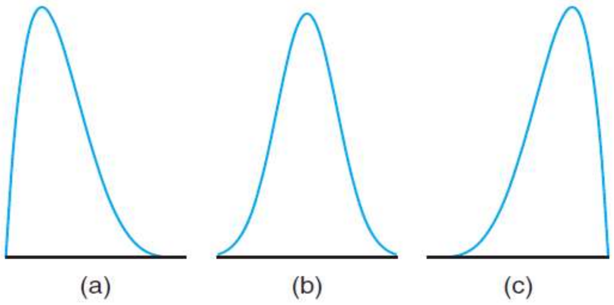


Figure 1.8: Skewness of data.

- A distribution is said to be symmetric if it can be folded along a vertical axis so that the two sides coincide.

- The distribution illustrated in Figure 2(a) is said to be skewed to the right since it has a long right tail and a much shorter left tail. In Figure 2(c) is said to be skewed to the left since it has a long left tail and a much shorter right tail.

Summarizing Data

Probability distributions:

- Show much more about a population than just the mean and standard deviation.
- A distribution may be symmetric, or may be skewed to the right or the left.
- The tail of a distribution shows the distance from the mean of the outlying points (for example, the 95th percentile point).

See You Next Lesson

Have a good week. Best Wishes…

See you ☺