Applied Statistics and Queuing Theory
Course No: CSE 3109

Introduction

Statistics and its importance

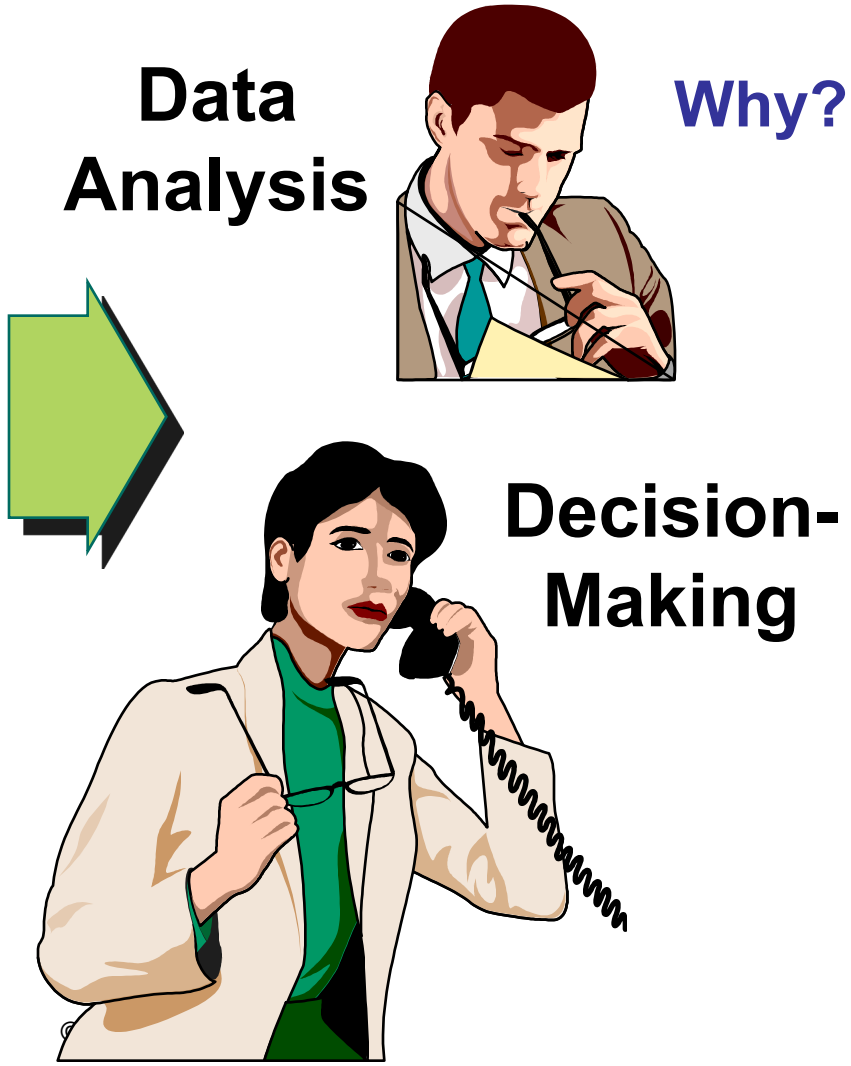
- **Statistics** is the study of the collection, organization, analysis, interpretation, and presentation of data.

or

- **Statistics:** The science of collecting, describing, and interpreting data.
- A general process of investigation consists of the steps:
  1. Research question identification.
  2. Data collection on the topic.
  3. Data analysis.
  4. Derive a conclusion based on the analysis of data.

Statistics and its importance

1. **Collecting Data**
   e.g., Survey
2. **Presenting Data**
   e.g., Charts & Tables
3. **Characterizing Data**
   e.g., Average



Statistics and its importance

Area we use statistics:

- Business and Industry
  - Statistics to Start or start a Business
  - Statistics to manufacturing
  - Statistics to marketing
  - Statistics to Engineering
  - Statistical Computing
- Health and Medicine
- Learning
  - Statistic for teachers
  - Result

Statistics and its importance

Area we use statistics (cont.):

- Research
- Social Statistics
  - Child-bearing, Child and elderly populations , Population
    - ✔ health, nutrition and educational level in country.
    - ✔ to identify the strength of working people.
    - ✔ to planning the future
  - Housing , Human settlements
    - ✔ identify problems in housing planning.
    - ✔ to settle the problems in slums

Statistics and its importance

Area we use statistics (cont.):

- Social Statistics
  - Education, Literacy
    - ✔ study about the currant education system in country.
    - ✔ develop the subject planning.
    - ✔ future employment planning
  - Health
    - ✔ to provide health facilities
  - Income and economic activity , Unemployment
    - ✔ to understand about savings and investment
    - ✔ introduce future investing systems
- Natural Resources

Population

- A **Population** is the collection or set of all items or individuals of interest

Or

- **Population** is the entire set of individuals or objects having some common characteristics selected for a research study
  - Examples: All likely voters in the next election
    All patients admitted to ICU
    All tax receipts over this year
- Two kinds of populations: **finite** or **infinite**.

Sample

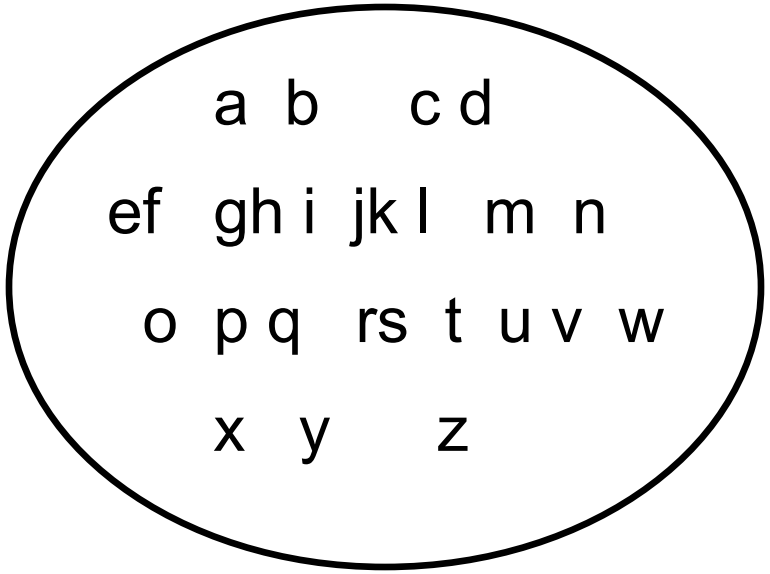
- A **Sample** is a subset of the population

Or

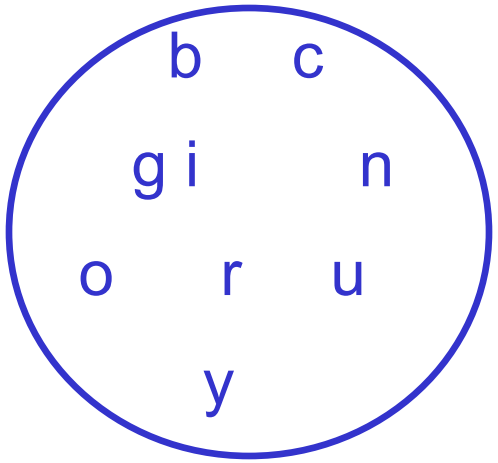
- **Sample** is the selected group of people or elements from which data are collected for a study
- Sample-
  - It is a unit that is selected from population
  - Represents the whole population
  - Purpose to draw the inference
  - Examples: 1000 voters selected at random for interview
    A few patients selected of heart disease
    Random receipts selected for audit
- **Sampling Frame** – Listing of population from which a sample is chosen

Population vs. Sample

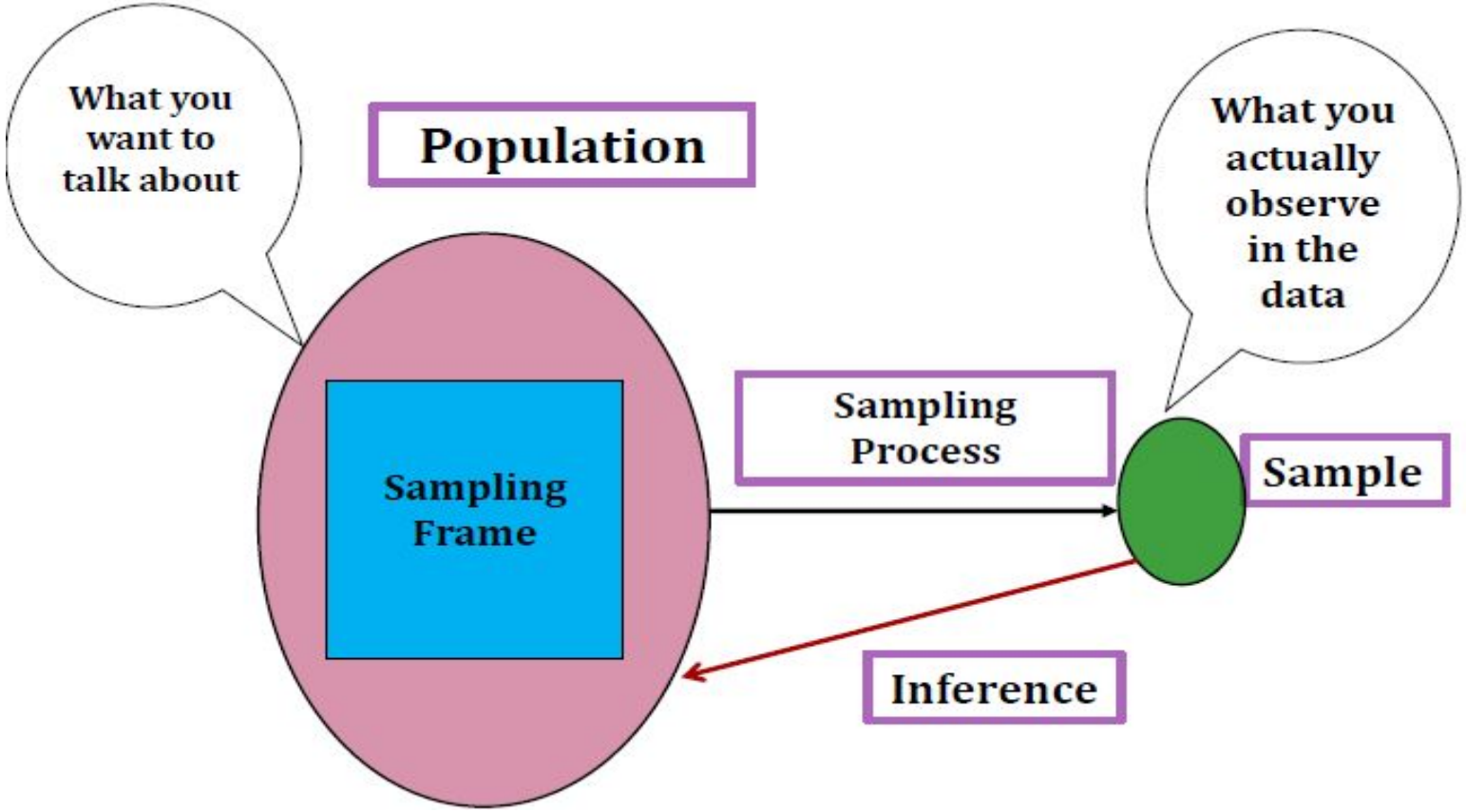
Population



Sample



Population vs. Sample



Sample

- **What is Good Sample?**

  The sample must be:

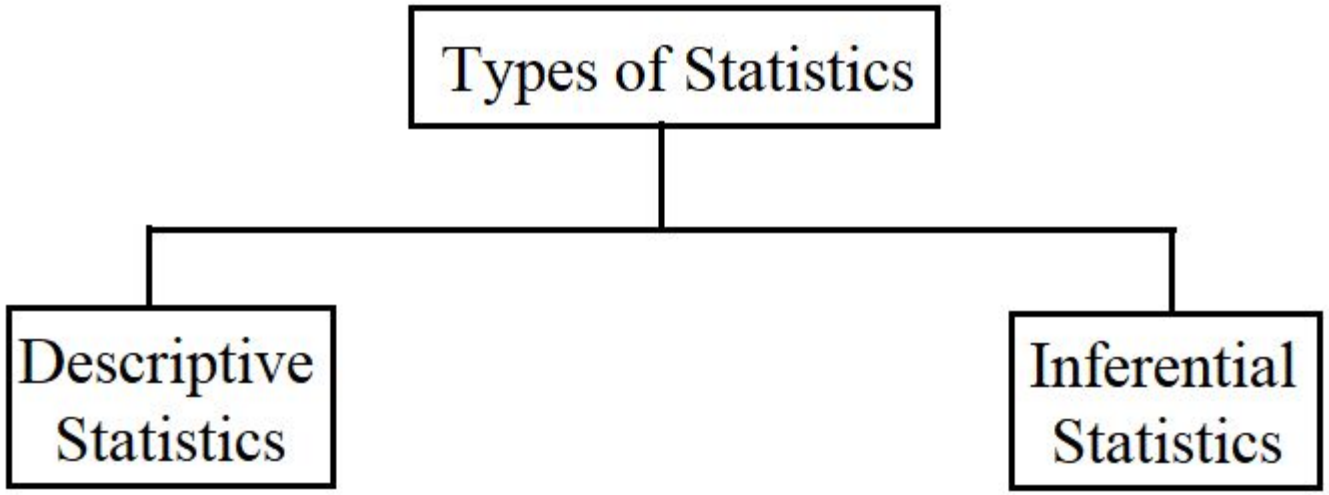
  1. representative of the population;
  2. appropriately sized (the larger the better);
  3. unbiased;
  4. random (selections occur by chance);

- **Merits of Sampling**
  - Size of population
  - Fund required for the study
  - Facilities
  - Time

Why Sample?

- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.

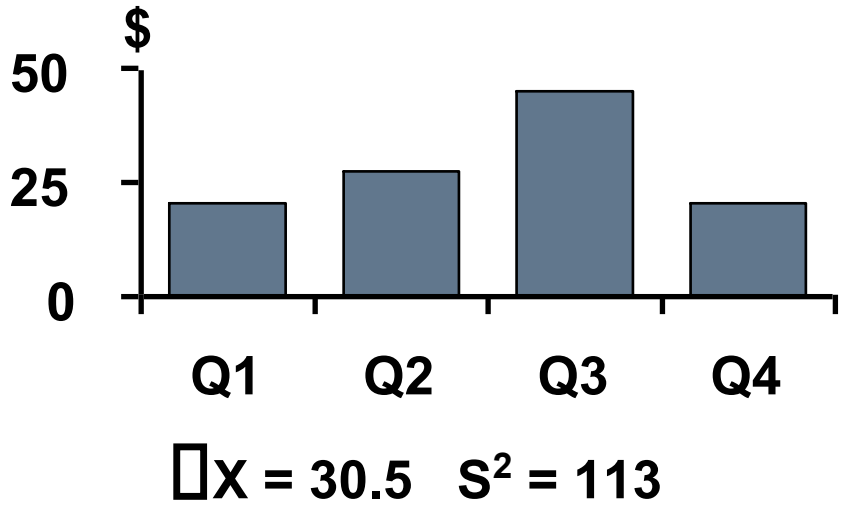
Types of Statistics



Types of Statistics (Cont'd)

Two types of statistics:

1. **Descriptive statistics**
   - Collecting, presenting, and analyzing data
   - Involves
     - Collecting Data
     - Presenting Data
     - Analyzing Data
   - Purpose
     - Describe Data



Types of Statistics (Cont'd)

Two types of statistics (cont.):

2. **Inferential statistics**
   - Drawing conclusions and/or making decisions concerning a population based only on sample data
   - Involves
     - Estimation
     - Hypothesis Testing
   - Purpose
     - Make decisions about population characteristics

Variable

In statistics, a **variable** has two defining characteristics:

- A variable is an attribute that describes a person, place, thing, or idea.
- The value of the variable can "vary" from one entity to another.

Definition

A variable is a characteristic, often but not always quantitatively measured, containing two or more values or categories that can vary from person to person, object to object or from phenomenon to phenomenon.

Constant

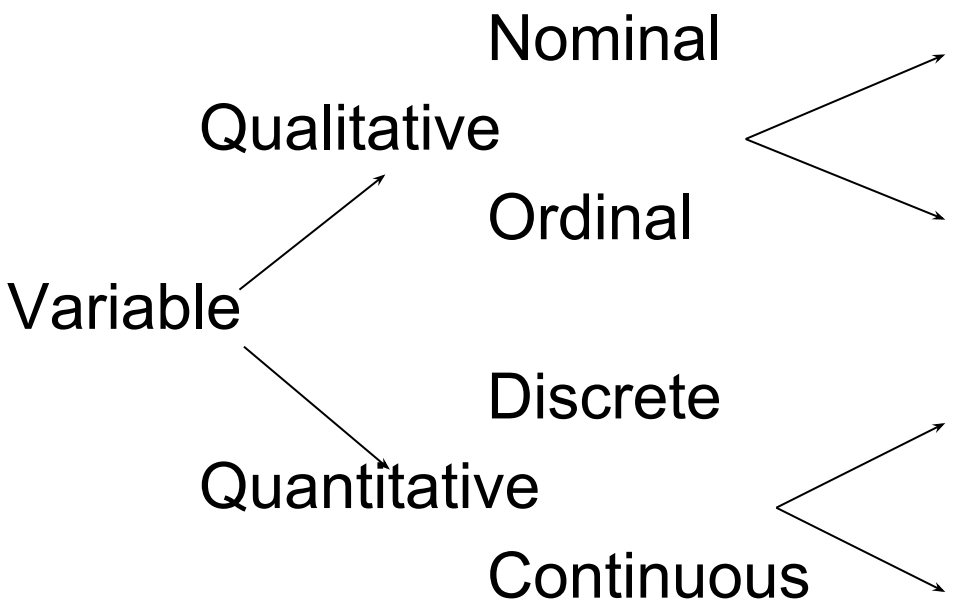
- A logical opposite of a variable is a constant.
- A constant is a particular type of variable, which does not vary from one member of a group to another

Definition

The term constant refers to a property whereby the members of a group or category remain fixed and do not one from another.

Variable and Constants

Variable Types:



Variable and Constants

Variable Types (cont.):

- **Quantitative/Numerical variable:**
  - A variable that quantifies an element of a population.
  - accepts numerical values.
  - Arithmetic operations like addition, subtraction, average, etc are meaningful.
- **Qualitative/Categorical Variable:**
  - A variable that categorizes or describes an element of a population.
  - no arithmetic operation
  - categories may be represented by numbers Like, male = 0, female = 1.

Variable and Constants

Variable Types (cont.):

- **Example:**
  - The residence hall for each student in a statistics class. (Categorical)
  - The amount of gasoline pumped by the next 10 customers at the local Unimart. (Numerical)
  - The amount of radon in the basement of each of 25 homes in a new development. (Numerical)
  - The color of the baseball cap worn by each of 20 students. (Categorical)
  - The length of time to complete a mathematics homework assignment. (Numerical)

Variable and Constants

Numerical/Quantitative Variable Types:

- **Discrete Variable:**
  - possible to count/enumerate all possible values e.g. number of rooms in a house.
  - in general, countable data is an example of discrete variable e.g. population in each division in Bangladesh.
  - non-negative whole numbers.
- **Continuous Variable:**
  - A quantitative variable that can assume an uncountable number of values
  - are usually associated with measurements e.g. height.
  - can accept any number of infinite values within a given range.

Variable and Constants

Qualitative/Categorical variable Types:

- **Nominal Variable:**
  - categories do not have an inherent(general) ordering.
  - e.g. do you prefer to write early in the morning or before going to bed at night? The answers can be {morning, night}.
- **Ordinal Variable:**
  - categories have an inherent ordering.
  - e.g. how satisfied you are with a customer service? The answers can be {very satisfied, satisfied, neutral, unsatisfied}.

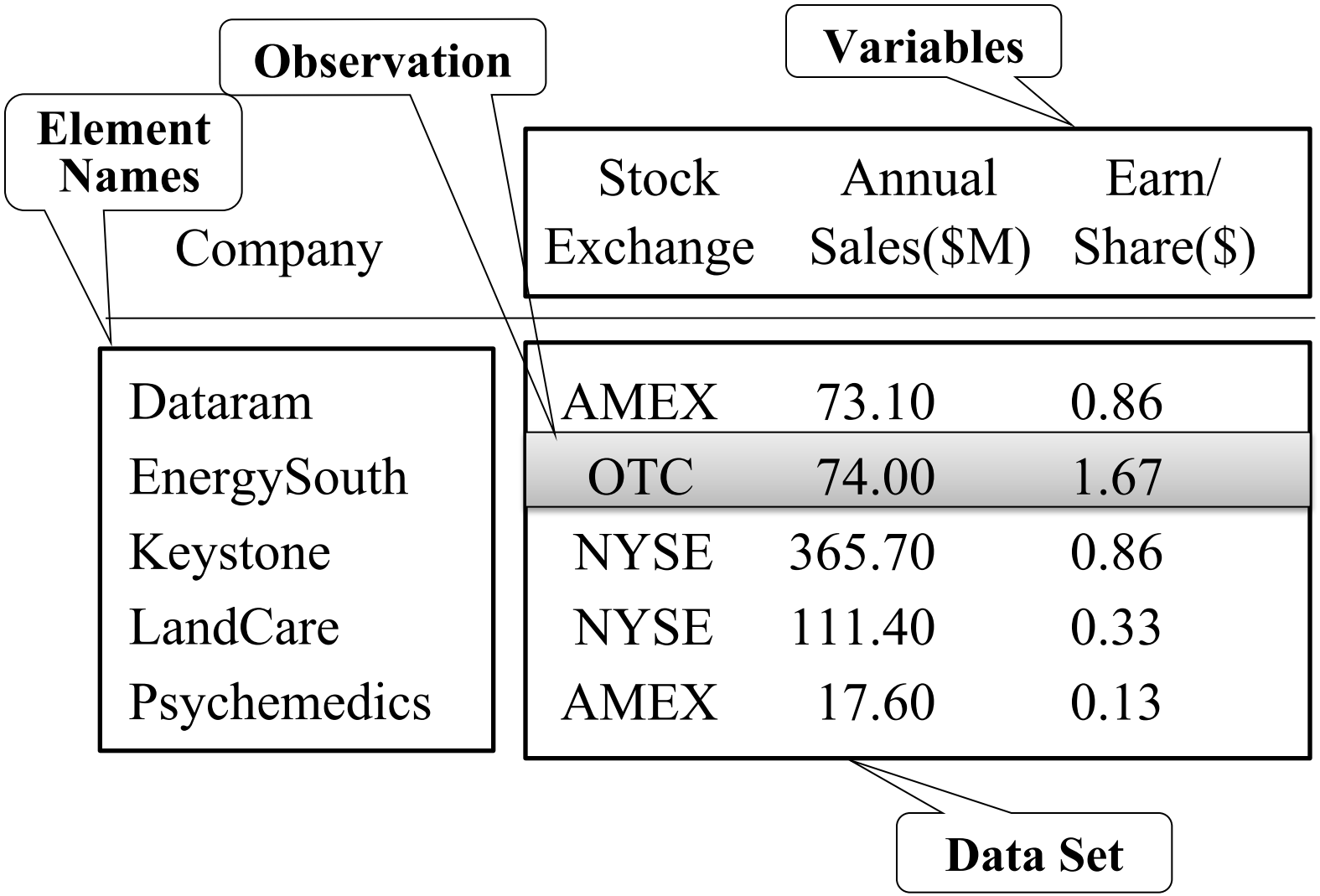
Statistical data

- **Data** are the facts and figures collected, summarized, analyzed, and interpreted.
  - IBM's sales revenue is $100; stock price $80
- The data collected in a particular study are referred to as the **data set.**
  - The sales revenue and stock price data for a number of firms including IBM, Dell, Apple, etc

Statistical data

- The **elements** are the entities on which data are collected.
  - IBM, Dell, Apple, etc. in the previous setting
- A **variable** is a characteristic about each individual elements.
  - Sales revenue, stock price (of a company)
- The set of measurements collected for a particular element is called an **observation**.
  - Sales revenue, stock price for 2003
- The total number of data values in a data set is the number of elements multiplied by the number of variables.

Statistical data



Nature of Data

- Continuous data – can include any value (i.e., real numbers)
  - e.g., 1, 1.43, and 3.1415926 are all acceptable values.
  - Geographic examples: distance, tree height, amount of precipitation, etc.
- Discrete data – only consists of discrete values, and the numbers in between those values are not defined (i.e., whole or integer numbers)
  - e.g., 1, 2, 3.
  - Geographic examples: # of vegetation types,

Nature of Data (Cont'd)

- Individual Data:
  - Individual data refers to raw, ungrouped data where each observation or data point is separate and distinct.
  - For example, if you have a list of ages of students in a class:

    18, 20, 22, 19, 21, 23, 20, 24, 19, 26
- Grouped Data:
  - Grouped data involves categorizing individual data points into intervals or classes and then counting the frequency of observations within each interval.

Age Interval	Frequency
15-19	3
20-24	6
25-29	1

Nature of Data (Cont'd)

- The scale determines the amount of information contained in the data.
- The scale indicates the data summarization and statistical analyses that are most appropriate
- Scales of measurement include:
  - Nominal
  - Ordinal
  - Interval
  - Ratio

Nature of Data (Cont'd)

- Nominal scale data are labels or names used to identify an attribute of the element
  - Example: Students of a university are classified by the dorm that they live in using a nonnumeric label such as Farley, Keenan, Zahm, Breen-Phillips, and so on.

    A numeric code can be used for the school variable (e.g. 1: Farley, 2: Keenan, 3: Zahm, and so on).
- Ordinal scale data have the properties of nominal data and the order or rank of the data is meaningful.
  - star-system restaurant rankings

    5 stars > 4 stars, 4 stars > 3 stars, 5 stars > 2 stars

Nature of Data (Cont'd)

Interval Scale Data:

- Interval data are measured and have constant, equal distances between values, but the zero point is arbitrary.
- There is no absolute zero.
- The zero isn't meaningful, it doesn't mean a true absence of something.

Example:

The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.

Nature of Data (Cont'd)

Ratio Scale Data:

- A *ratio* variable, has all the properties of an interval variable, and also has a clear definition of 0.0.
- Ratio scales have an absolute zero
- When the variable equals 0.0, there is none of that variable.

Examples

Variables like height, weight, enzyme activity are ratio variables.

Interval Scales vs. Ratio Scales

- Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no heat'.
- However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no heat'.
- Another counter example is pH. It is not a ratio variable, as pH=0 just means 1 molar of H+. and the definition of molar is fairly arbitrary. A pH of 0.0 does not mean 'no acidity' (quite the opposite!).
- When working with ratio variables, but not interval variables, you can look at the ratio of two measurements.
- A weight of 4 grams is twice a weight of 2 grams, because weight is a ratio variable. A temperature of 100 degrees C is not twice as hot as 50 degrees C, because temperature C is not a ratio variable.
- A pH of 3 is not twice as acidic as a pH of 6, because pH is not a ratio variable.

Univariate vs. Multivariate Data

Statistical data are often classified according to the number of variables being studied.

Univariate data

When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of high school students. Since we are only working with one variable (weight), we would be working with univariate data.

Multivariate data.

When we conduct a study that examines the relationship among more than two variables, we are working with multivariate data. Suppose we conducted a study to see if there were a relationship among the height, weight, and age of high school students. Since we are working with three variables (height , weight, age), we would be working with multivariate data.

Data Collection

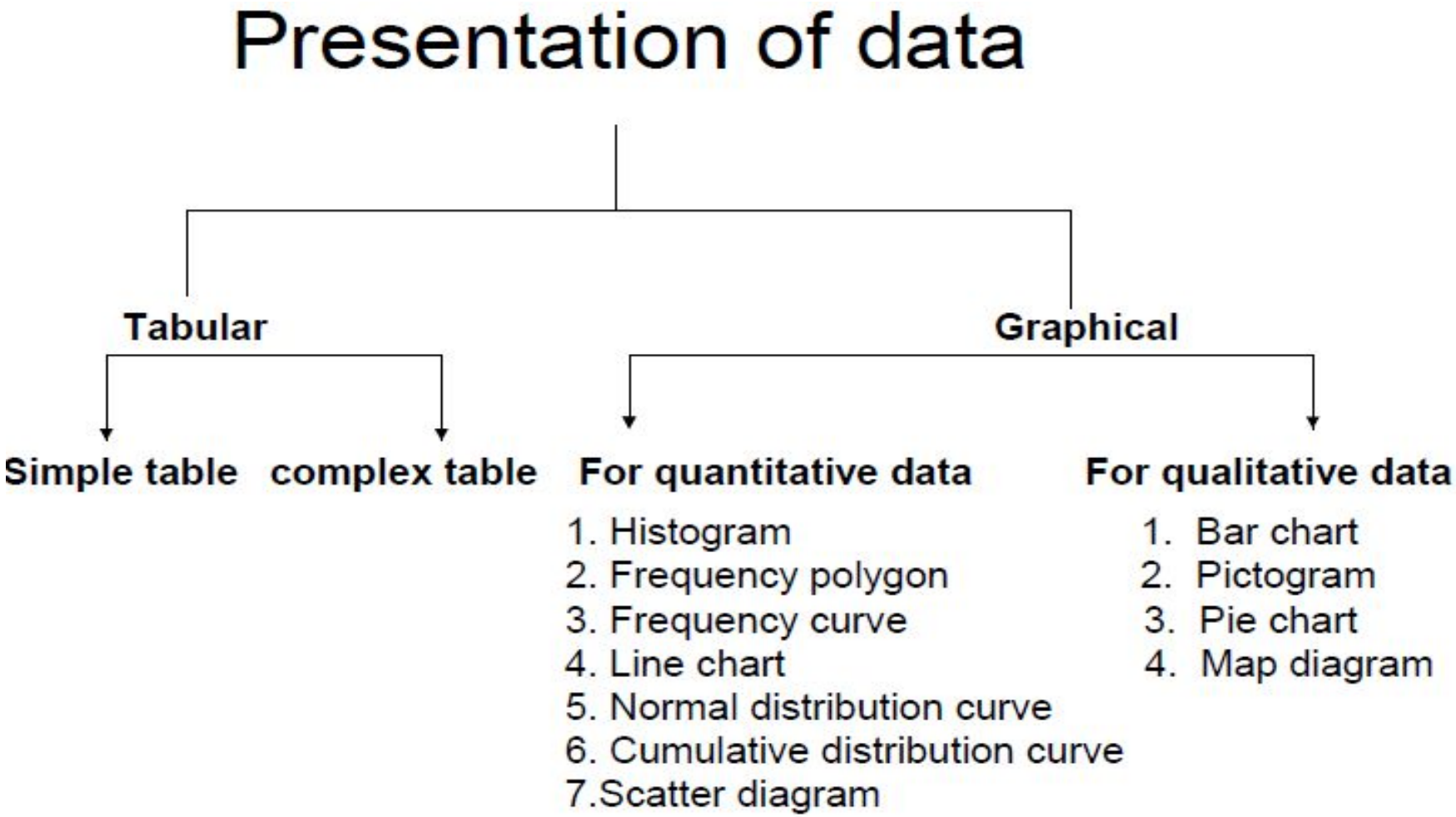
Obtaining Data:

- Data from a published source
  - book, journal, newspaper, Web site
- Data from a designed experiment
  - researcher exerts strict control over units
- Data from a survey
  - a group of people are surveyed and their responses are recorded
- Data collected observationally
  - units are observed in natural setting and variables of interest are recorded

Data Presentation

- An effective presentation of the data often quickly reveals important features such as
  - their range,
  - degree of symmetry,
  - how concentrated or spread out they are,
  - where they are concentrated, and so on.

Data Presentation



Tabular Presentation

- Tables are the devices, that are used to present the data in a simple form. It is probably the first step before the data is used for analysis or interpretation.
- General principals of designing tables
  - The tables should be numbered e.g table 1, table 2 etc.
  - A title must be given to each table.
  - The headings of columns or rows should be clear and concise.
  - The data must be presented according to size or importance chronologically, alphabetically, or geographically.
  - If percentages or averages are to be compared, they should be placed as close as possible.
  - No table should be too large

Tabular Presentation

- **Types of tables**
  - **Simple tables:** Measurements of single set are presented
  - **Complex tables:** Measurements of multiple sets are presented

Table 2.1 A Frequency Table of Sick Leave Data

Value	Frequency	Value	Frequency
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

Simple table

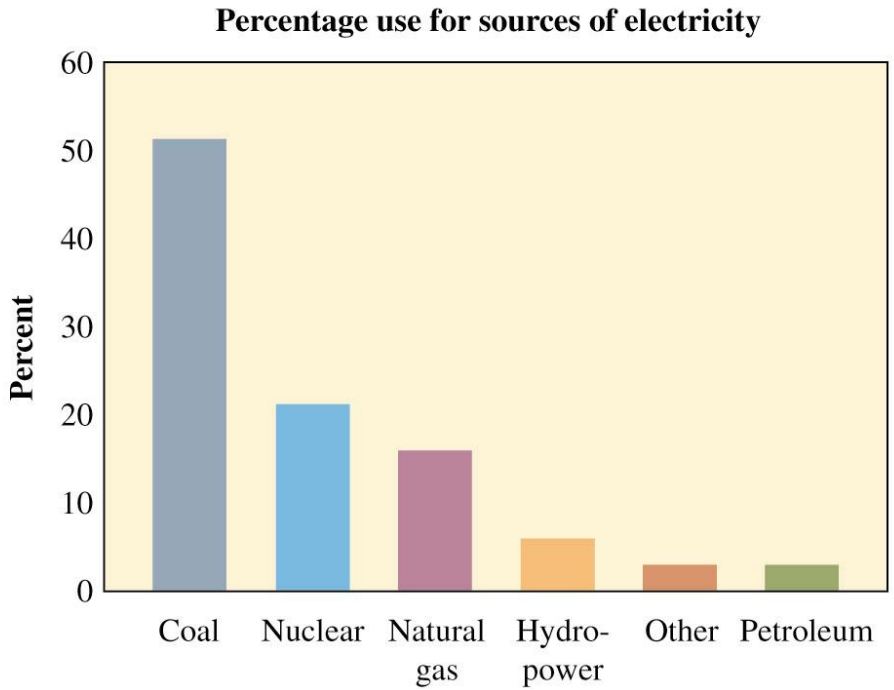
Region	Apple	Orange
North America		
2002	$10	$12
2003	$11	$13
2004	$12	$14
2005	$11	$13
2006	$10	$14
Average	$10.80	$13.20
Europe		
2002	$11	$13
2003	$11	$14
2004	$13	$13
2005	$12	$14
2006	$11	$15
Average	$11.60	$13.80
Overall Average	$11.20	$13.50

Complex table

Bar Graphs

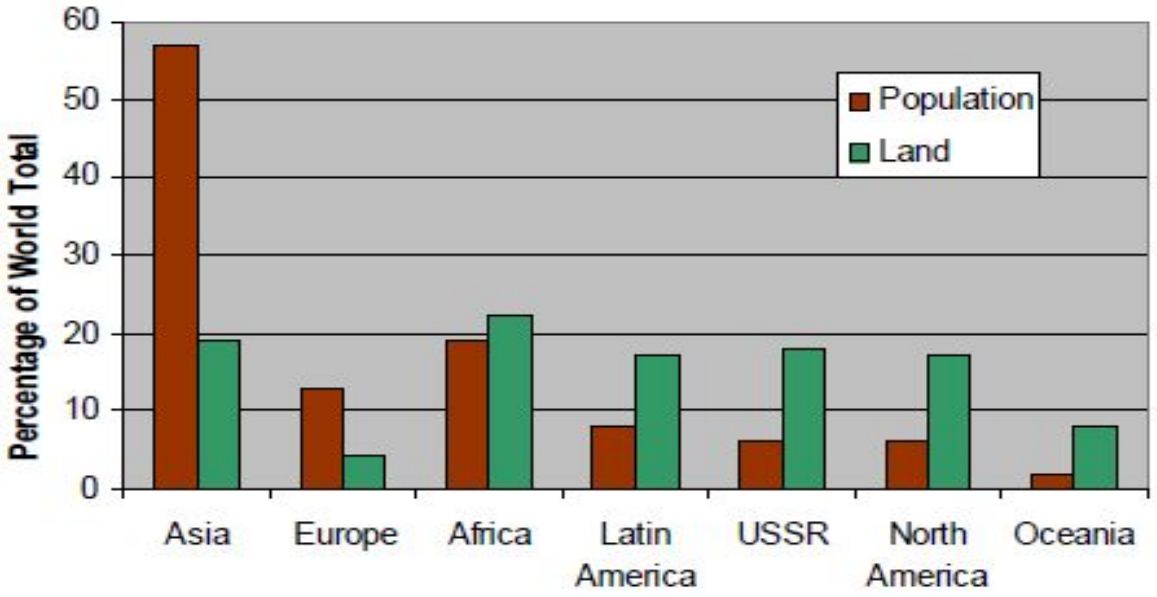
- Summarizes categorical variable
- Vertical bars for each category
- Height of each bar represents either counts or percentages
- Easier to compare categories with bar graph than with pie chart
- Called Pareto Charts when ordered from tallest to shortest

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	**100**



Multiple Bar Graphs

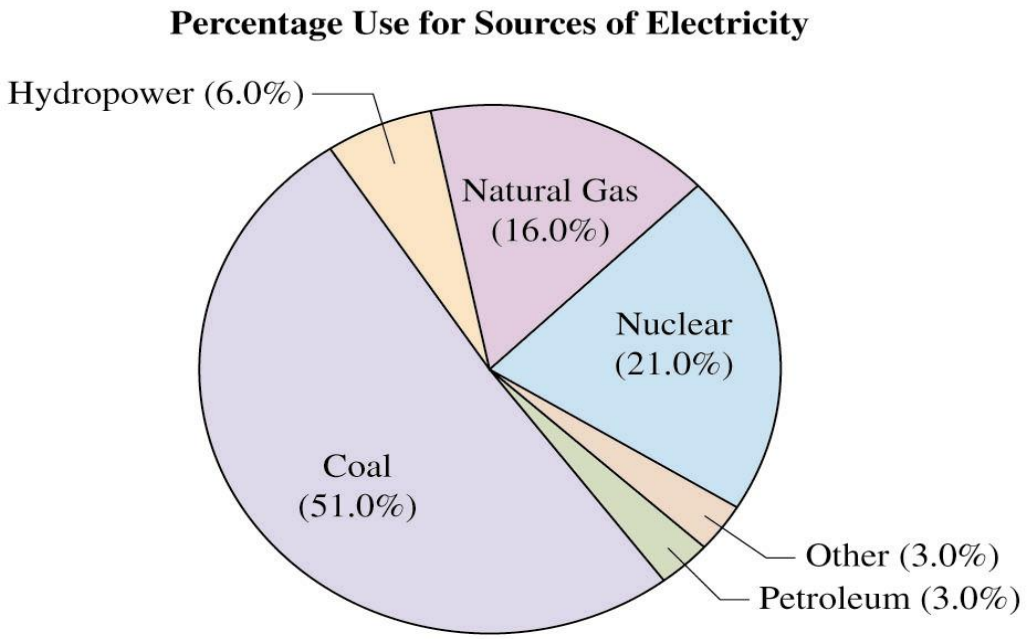
- Also called compound bar charts
- More then one sub-attribute of variable can be expressed
- Used for compare data



Pie Charts

- Summarize categorical variable
- Drawn as circle where each category is a slice
- The size of each slice is proportional to the percentage in that category

Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	**100**



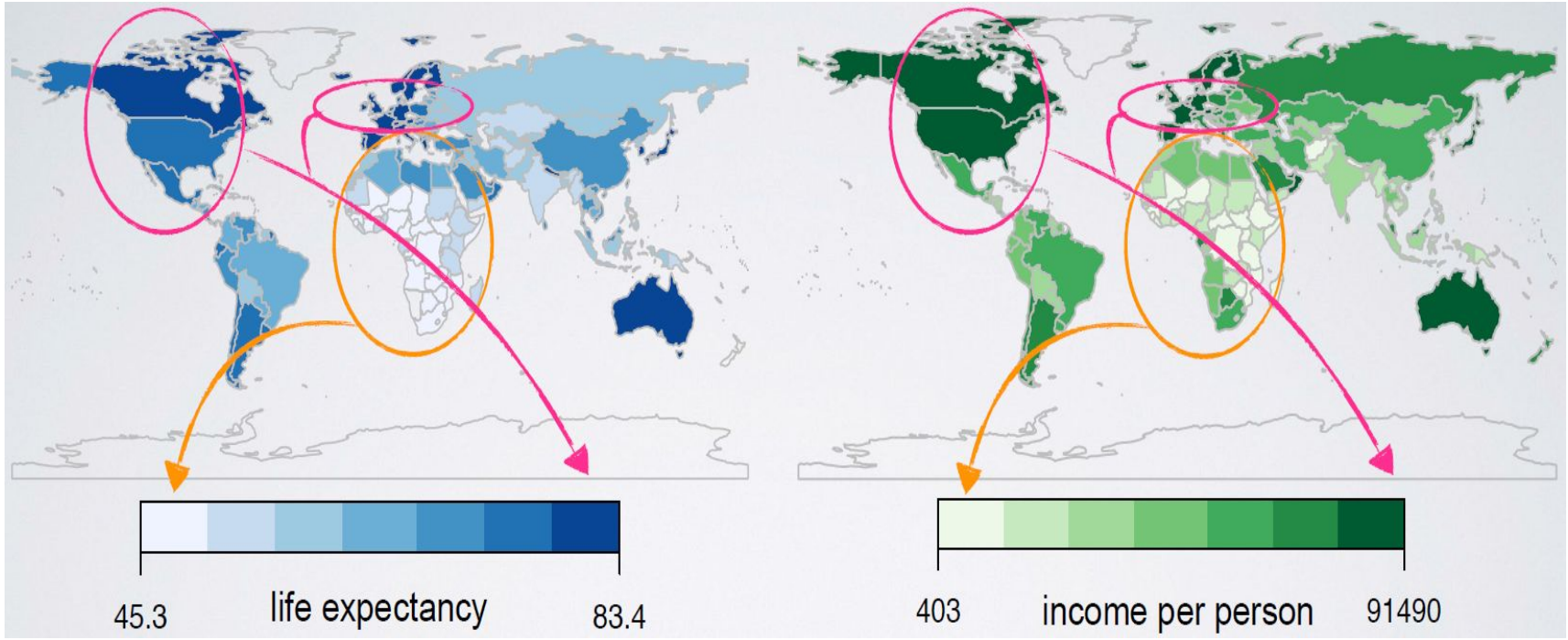
Pictogram

- Small pictures or symbols are used to present the data.
- Fraction of the picture can be used to represent numbers smaller than the value of whole symbol

FRUIT	NUMBER OF CHILDREN WHO CHOSE IT
PEAR	
WATERMELON	
ORANGE	
APPLE	
BANANA	

Map Diagram

- When statistical data refers to geographic or administrative areas, it is presented either as statistical map or dot map.
- useful for highlighting the spatial distribution.



life expectancy and income are lower in Africa but both are higher in Europe.

Frequency Distribution

- **Frequency Distribution:** A listing, often expressed in chart form, that pairs each value of a variable with its frequency

- **Ungrouped Frequency Distribution:** Each value of x in the distribution stands alone

- **Grouped Frequency Distribution:** Group the values into a set of classes

Ungrouped Frequency Distribution

- The following data represent the number of days of sick leave taken by each of 50 workers of a given company over the last 6 weeks:

  2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1,8, 9, 7, 0,1, 7,
  2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

- Frequency tables

Table 2.1 A Frequency Table of Sick Leave Data

Value	Frequency	Value	Frequency
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

Example 2.1

- Use Table 2.1 to answer the following questions:

(a) How many workers had at least 1 day of sick leave?

(b) How many workers had between 3 and 5 days of sick leave?

(c) How many workers had more than 5 days of sick leave?

- **Solution**

(a) Since 12 of the 50 workers had no days of sick leave, the answer is $50 - 12 = 38$.

(b) The answer is the sum of the frequencies for values 3, 4, and 5; that is, $4 + 5 + 8 = 17$.

Table 2.1 A Frequency Table of Sick Leave Data

Value	Frequency	Value	Frequency
0	12	5	8
1	8	6	0
2	5	7	5
3	4	8	2
4	5	9	1

(c) The answer is the sum of the frequencies for the values 6, 7, 8, and 9. Therefore, the answer is $0 + 5 + 2 + 1 = 8$.

Grouped Frequency Distribution

- In the frequency distribution table, the data is first split up into convenient groups (class interval) and the number of items (frequency) which occur in each group is shown in adjacent columns.
- Hence it is a table showing the frequency with which the values are distributed in different groups or classes with some defined characteristics.
- Group the values into a set of classes.
- A table that summarizes data by classes, or class intervals.
- The table may contain columns for class number, class interval, frequency, relative frequency, cumulative relative frequency, and class midpoint.

Grouped Frequency Distribution

- Example:

Table 2.2

Heights of 100 male students at XYZ University

Height (in)	Number of Students
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8
	Total 100

Frequency Distribution Table

- **Class Interval:** A symbol defining a class, such as 60–62 in Table 2, is called a class interval.
- **Class Limits:** The end numbers, 60 and 62, are called class limits; the smaller number (60) is the lower class limit, and the larger number (62) is the upper class limit.
- **Class Boundaries:** If heights are recorded to the nearest inch, the class interval 60–62 theoretically includes all measurements from 59.5000 to 62.5000 in. These numbers, indicated briefly by the exact numbers 59.5 and 62.5, are called class boundaries, or true class limits; the smaller number (59.5) is the lower class boundary, and the larger number (62.5) is the upper class boundary

Frequency Distribution Table

- **The Size, or Width, of a Class Interval:** The size, or width, of a class interval is the difference between the lower and upper class boundaries and is also referred to as the class width, class size, or class length.
- If all class intervals of a frequency distribution have equal widths, this common width is denoted by c. In such case c is equal to the difference between two successive lower class limits or two successive upper class limits.
- For the data of Table 2, for example, the class interval is

$$c = 62.5 - 59.5 = 63 - 60 = 65 - 62 = 3$$

- The class mark is the midpoint of the class interval and is obtained by adding the lower and upper class limits and dividing by 2. Thus the class mark of the interval 60–62 is (60+62)/2 = 61.

Frequency Distribution Table

Rules for construction of frequency table

- Determine the highest(H) and lowest(L) numbers in the raw data and thus find the range. Range = $H - L$
- Divide the range into a convenient number of class intervals having the same size. If this is not feasible, use class intervals of different sizes or open class intervals. The number of class intervals is usually between 5 and 20, depending on the data.
- Pick a starting point a little smaller than L. Count from L by the width to obtain the class limits. Observations that fall on class limits are placed into the class interval to the right.
- Left-end inclusion convention
  - a class interval contains its left-end but not its right-end boundary point.
  - for instance the class interval 20–30 contains all values that are both greater than or equal to 20 and less than 30
- Determine the number of observations falling into each class interval; that is, find the class frequencies. This is best done by using a tally, or score sheet.

Frequency Distribution Table

- Example: In the following table the weights of 40 male students at State University are recorded to the nearest pound. Construct a frequency distribution.

138 164 150 132 144 125 149 157

146 158 140 147 136 148 152 144

168 126 138 176 163 119 154 165

146 173 142 147 135 153 140 135

161 145 135 142 150 156 145 128

Frequency Distribution Table

Weight (lb)	Tally	Frequency
118–122	/	1
123–127	//	2
128–132	//	2
133–137	////	4
138–142	///// /	6
143–147	///// ///	8
148–152	/////	5
153–157	////	4
158–162	//	2
163–167	///	3
168–172	/	1
173–177	//	2
	Total	40

Weight (lb)	Tally	Frequency
118–126	///	3
127–135	/////	5
136–144	///// ////	9
145–153	///// ///// //	12
154–162	/////	5
163–171	////	4
172–180	//	2
	Total	40

Example

✔ Example: The hemoglobin test, a blood test given to diabetics during their periodic checkups, indicates the level of control of blood sugar during the past two to three months. The data in the table below was obtained for 40 different diabetics at a university clinic that treats diabetic patients:

6.5	5.0	5.6	7.6	4.8	8.0	7.5	7.9	8.0	9.2
6.4	6.0	5.6	6.0	5.7	9.2	8.1	8.0	6.5	6.6
5.0	8.0	6.5	6.1	6.4	6.6	7.2	5.9	4.0	5.7
7.9	6.0	5.6	6.0	6.2	7.7	6.7	7.7	8.2	9.0

1) Construct a grouped frequency distribution using the classes 3.7 - 4.7, 4.7 - 5.7, 5.7 - 6.7, etc.

2) Which class has the highest frequency?

Solutions

1)

Class Boundaries	Frequency *f*	Relative Frequency	Cumulative Rel. Frequency	Class Midpoint, *x*
3.7 - 4.7	1	0.025	0.025	4.2
4.7 - 5.7	6	0.150	0.175	5.2
5.7 - 6.7	16	0.400	0.575	6.2
6.7 - 7.7	4	0.100	0.675	7.2
7.7 - 8.7	10	0.250	0.925	8.2
8.7 - 9.7	3	0.075	1.000	9.2

2) The class 5.7 - 6.7 has the highest frequency. The frequency is 16 and the relative frequency is 0.40

Relative frequency = (frequency/N)

Cumulative Frequency Distribution

Cumulative Frequency Distribution: A frequency distribution that pairs cumulative frequencies with values of the variable

- The *cumulative frequency* for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values

- The *cumulative relative frequency* for any given class is the sum of the relative frequency for that class and the relative frequencies of all classes of smaller values

Example

- Example: A computer science aptitude test was given to 50 students. The table below summarizes the data:

Class Boundaries	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Rel. Frequency
0 up to 4	4	0.08	4	0.08
4 up to 8	8	0.16	12	0.24
8 up to 12	8	0.16	20	0.40
12 up to 16	20	0.40	40	0.80
16 up to 20	6	0.12	46	0.92
20 up to 24	3	0.06	49	0.98
24 up to 28	1	0.02	50	1.00

Ogive

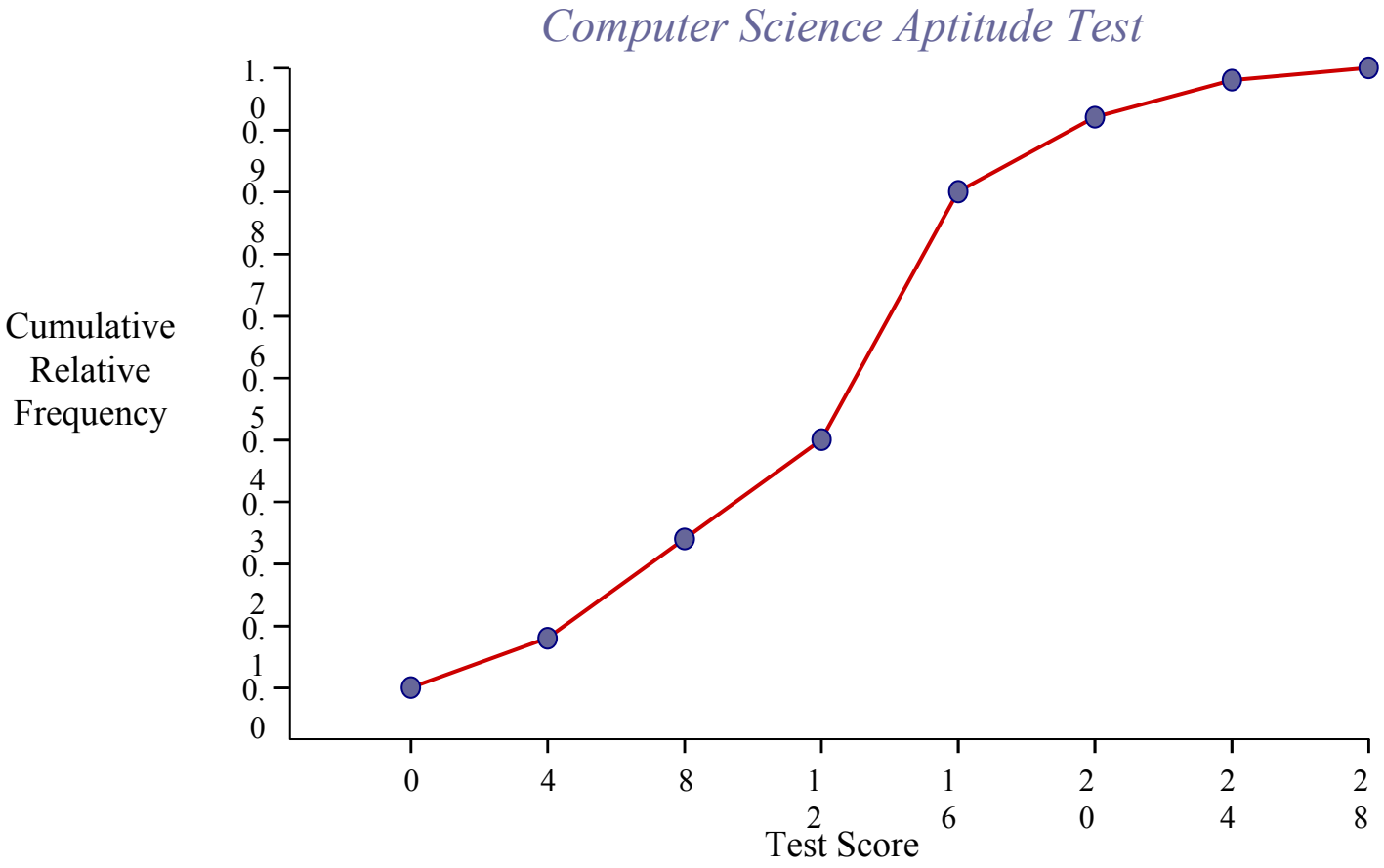
Ogive: A line graph of a cumulative frequency or cumulative relative frequency distribution. An ogive has the following components:

1. A title, which identifies the population or sample
2. A vertical scale, which identifies either the cumulative frequencies or the cumulative relative frequencies
3. A horizontal scale, which identifies the upper class boundaries. Until the upper boundary of a class has been reached, you cannot be sure you have accumulated all the data in the class. Therefore, the horizontal scale for an ogive is always based on the upper class boundaries.

Note: Every ogive starts on the left with a relative frequency of zero at the lower class boundary of the first class and ends on the right with a relative frequency of 100% at the upper class boundary of the last class.

Example

✔ Example: The graph below is an ogive using cumulative relative frequencies for the computer science aptitude data:



Class Boundaries	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Rel. Frequency
0 up to 4	4	0.08	4	0.08
4 up to 8	8	0.16	12	0.24
8 up to 12	8	0.16	20	0.40
12 up to 16	20	0.40	40	0.80
16 up to 20	6	0.12	46	0.92
20 up to 24	3	0.06	49	0.98
24 up to 28	1	0.02	50	1.00

Stem & Leaf Display

- Background:
  - The stem-and-leaf display has become very popular for summarizing numerical data
  - It is a combination of graphing and sorting
  - The actual data is part of the graph
  - Well-suited for computers

Stem-and-Leaf Display: Pictures the data of a sample using the actual digits that make up the data values. Each numerical data is divided into two parts: The leading digit(s) becomes the *stem*, and the trailing digit(s) becomes the *leaf*. The stems are located along the main axis, and a leaf for each piece of data is located so as to display the distribution of the data.

Example

- **Example:** A city police officer, using radar, checked the speed of cars as they were traveling down the main street in town. Construct a stem-and-leaf plot for this data:

41	31	33	35	36	37	39	49
33	19	26	27	24	32	40	
39	16	55	38	36			

Solution:

All the speeds are in the 10s, 20s, 30s, 40s, and 50s. Use the first digit of each speed as the stem and the second digit as the leaf. Draw a vertical line and list the stems, in order to the left of the line. Place each leaf on its stem: place the trailing digit on the right side of the vertical line opposite its corresponding leading digit.

Example

20 Speeds

```
---------------------------------------
1 | 6  9
2 | 4  6  7
3 | 1  2  3  3  5  6  6  7  8  9  9
4 | 0  1  9
5 | 5
---------------------------------------
```

- The speeds are centered around the 30s

Histogram

Histogram: A bar graph representing a frequency distribution of a quantitative variable. A histogram is made up of the following components:

1. A title, which identifies the population of interest
2. A vertical scale, which identifies the frequencies in the various classes
3. A horizontal scale, which identifies the variable x. Values for the class boundaries or class midpoints may be labeled along the x-axis. Use whichever method of labeling the axis best presents the variable.

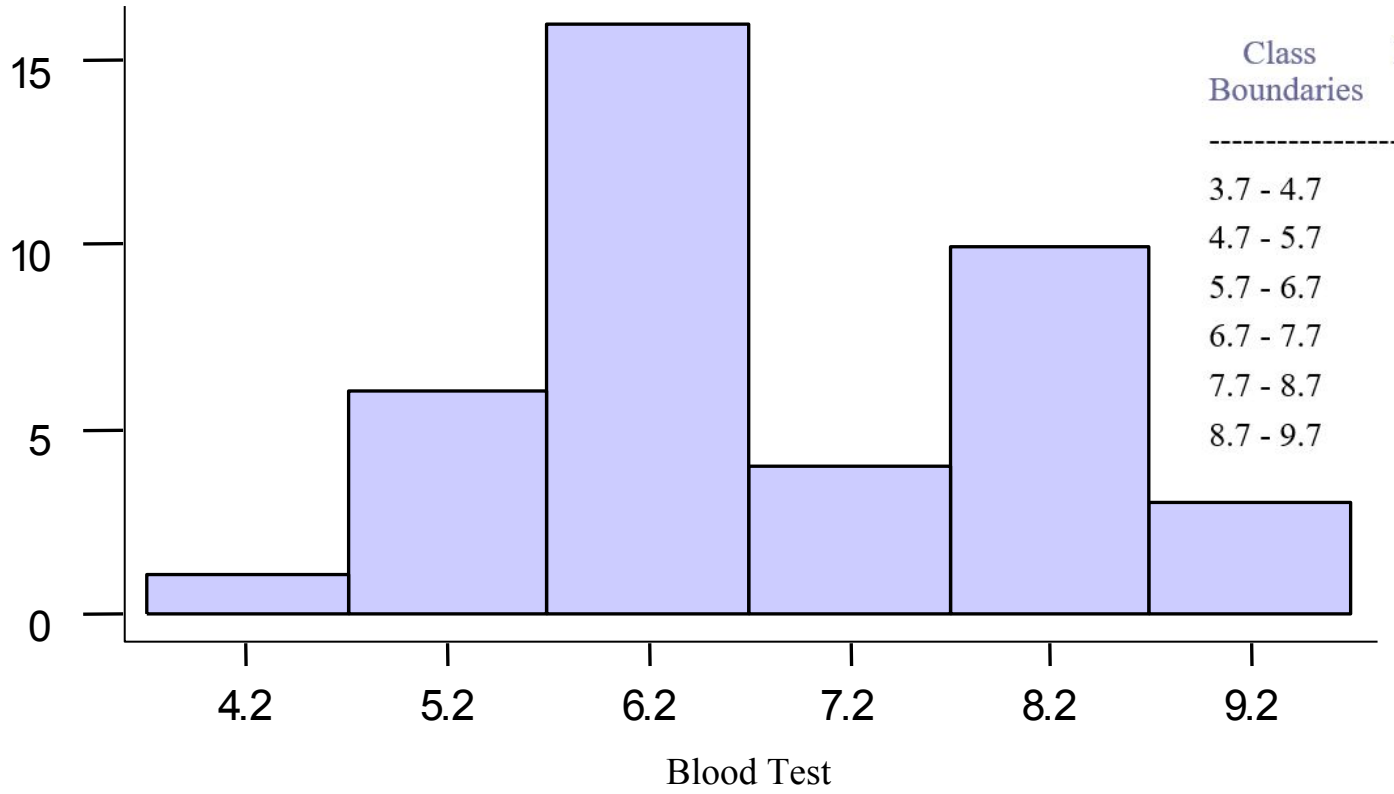
Notes:

- The relative frequency is sometimes used on the vertical scale
- It is possible to create a histogram based on class midpoints

Example

✔ Example: Construct a histogram for the blood test results given in the previous example

Solution:



Class Boundaries	Frequency f	Relative Frequency	Cumulative Rel. Frequency	Class Midpoint, x
3.7 - 4.7	1	0.025	0.025	4.2
4.7 - 5.7	6	0.150	0.175	5.2
5.7 - 6.7	16	0.400	0.575	6.2
6.7 - 7.7	4	0.100	0.675	7.2
7.7 - 8.7	10	0.250	0.925	8.2
8.7 - 9.7	3	0.075	1.000	9.2

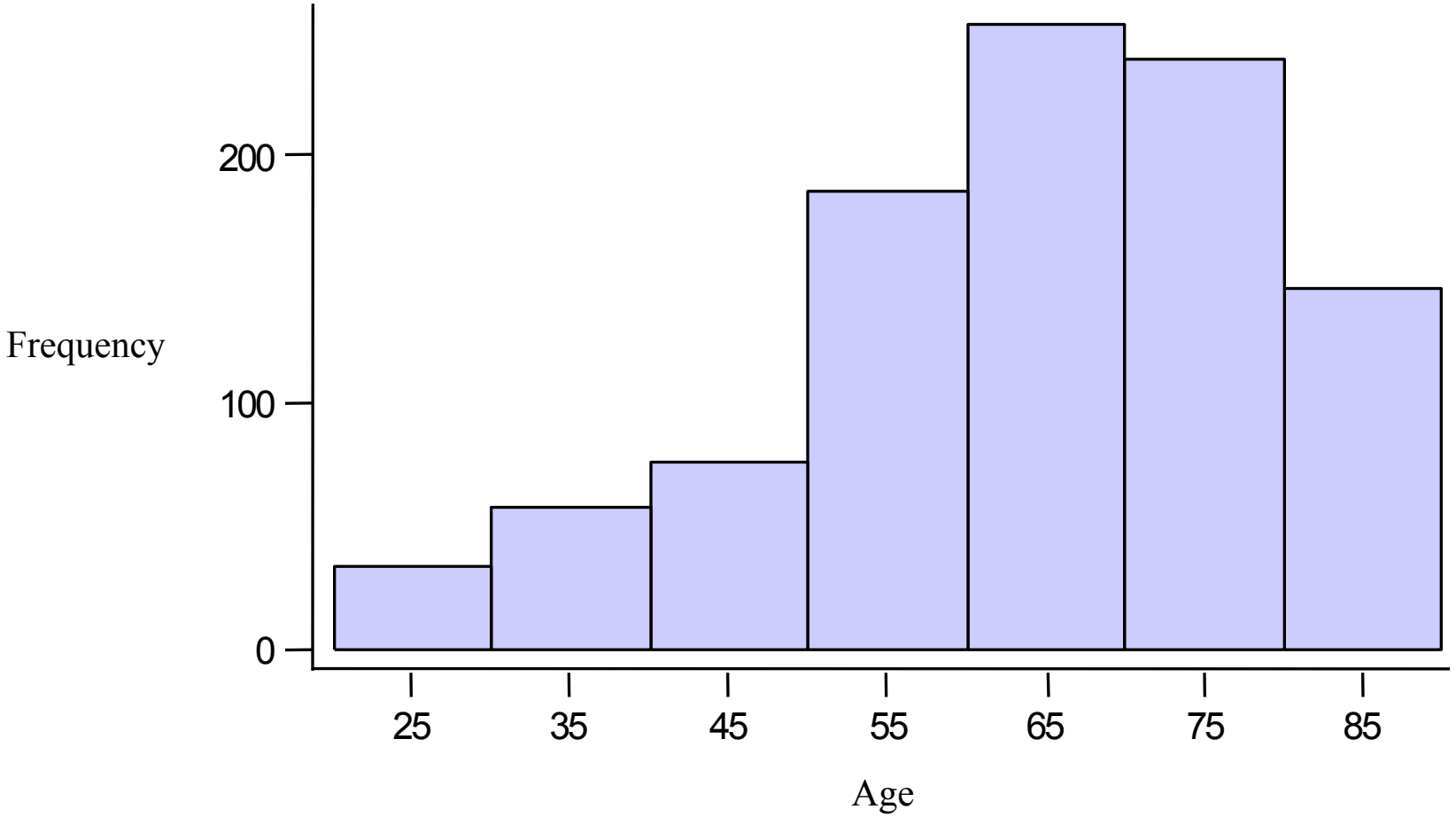
Example

- Example: A recent survey of Roman Catholic nuns summarized their ages in the table below. Construct a histogram for this age data:

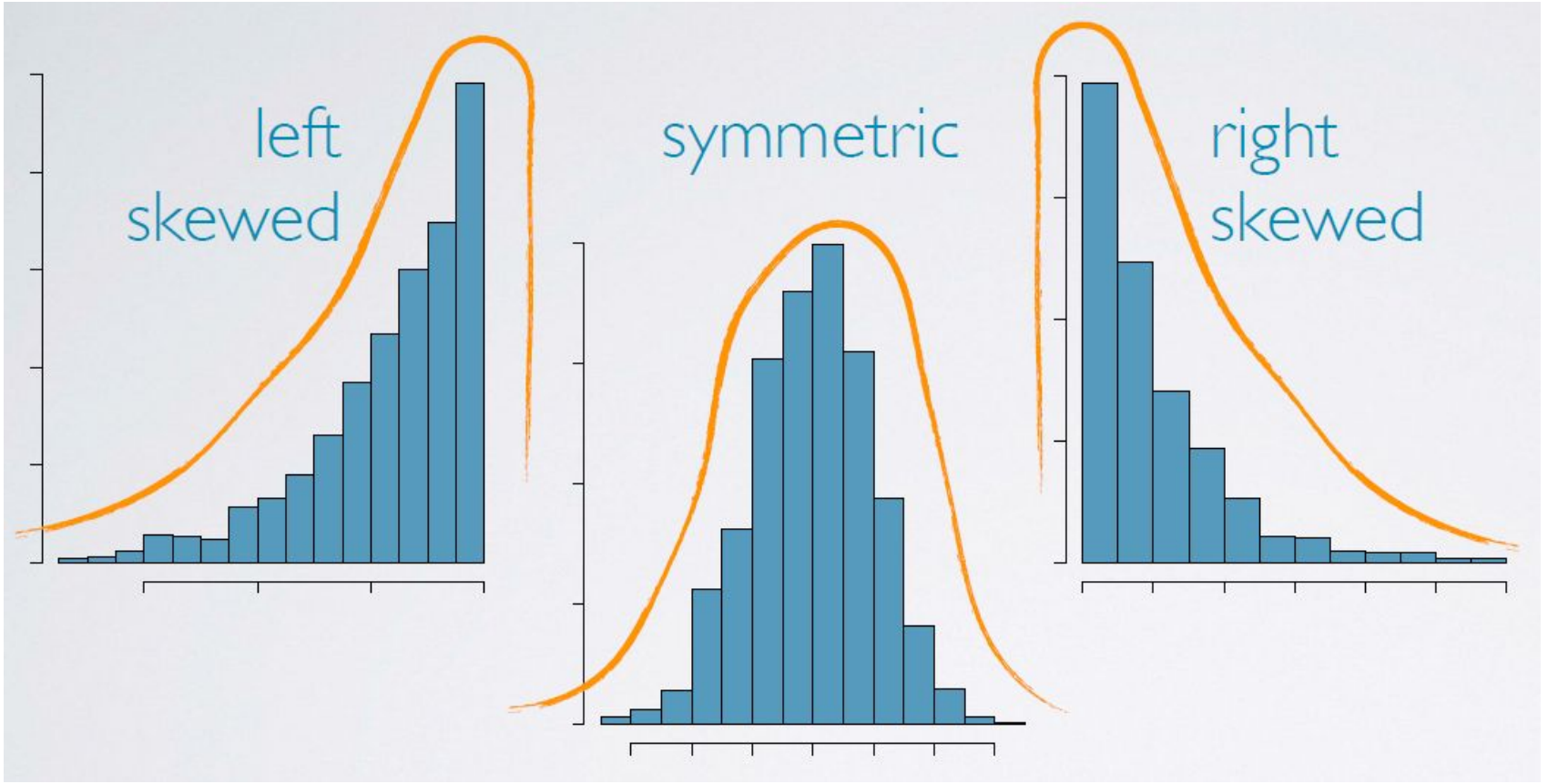
Age	Frequency	Class Midpoint
20 up to 30	34	25
30 up to 40	58	35
40 up to 50	76	45
50 up to 60	187	55
60 up to 70	254	65
70 up to 80	241	75
80 up to 90	147	85

Solution

Roman Catholic Nuns



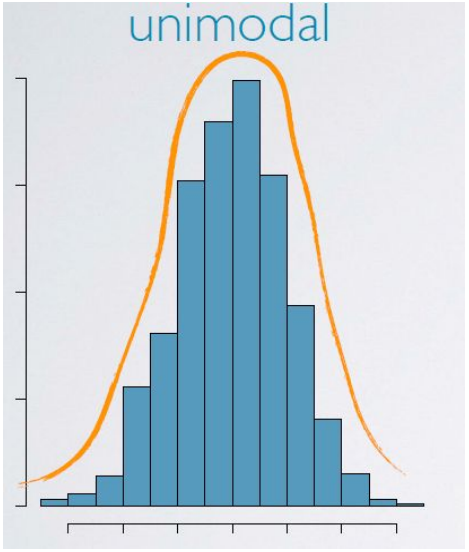
Skewness of Histogram



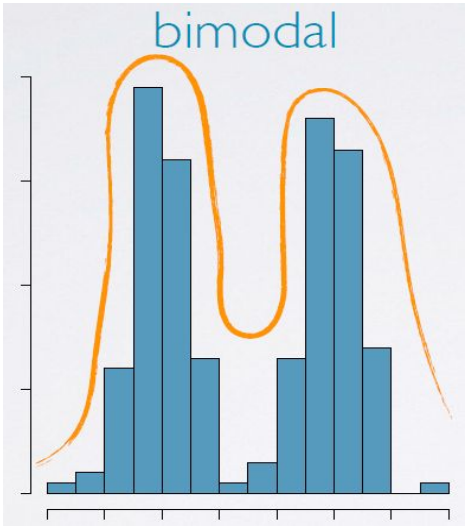
- distributions are said to be skewed to the side of the long tail.
- if no skewness is apparent, the distribution is said to be symmetric.

Modality of Histogram

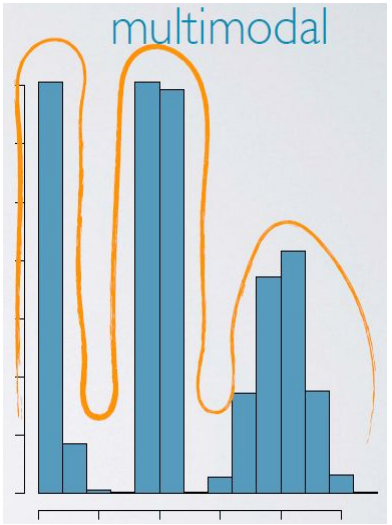
- Modality is also related to the shape of a histogram
- Refers to the number of prominent peaks



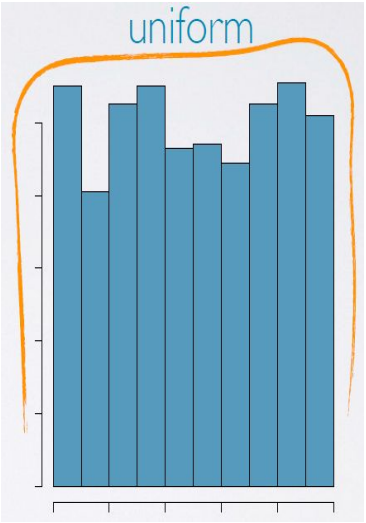
Uni-modal
one prominent peak



Bi-modal
two prominent peaks



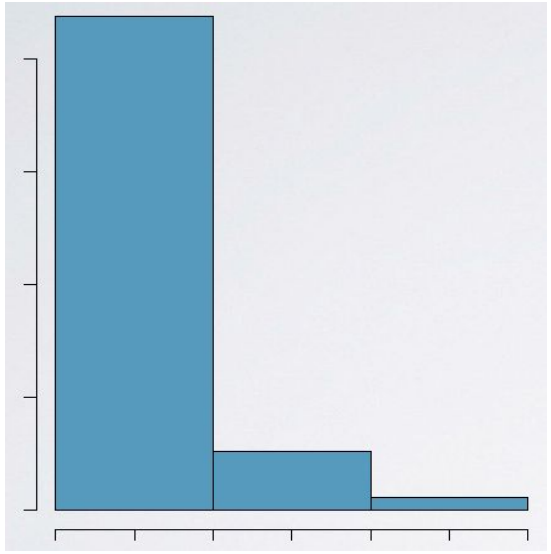
Multimodal
More than two prominent peaks



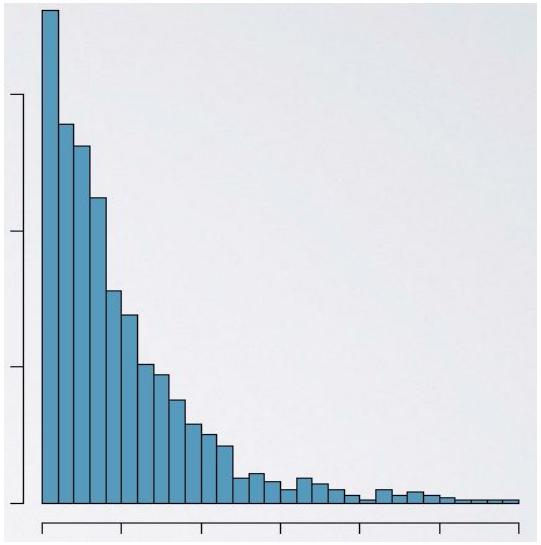
Uniform
no prominent peak

Bin Width of Histogram

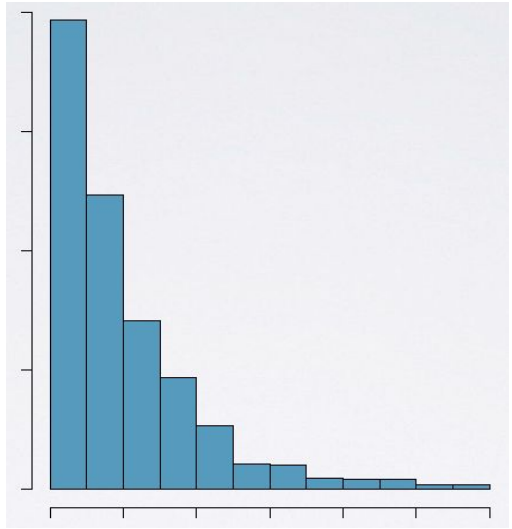
- The chosen bin width can alter the story that the histogram is telling.
- Too large bin width → we may lose interesting details.
- Too narrow bin width → difficult to get the overall picture of the distribution.
- Ideal bin width depends on the data being analyzed



too wide



too narrow



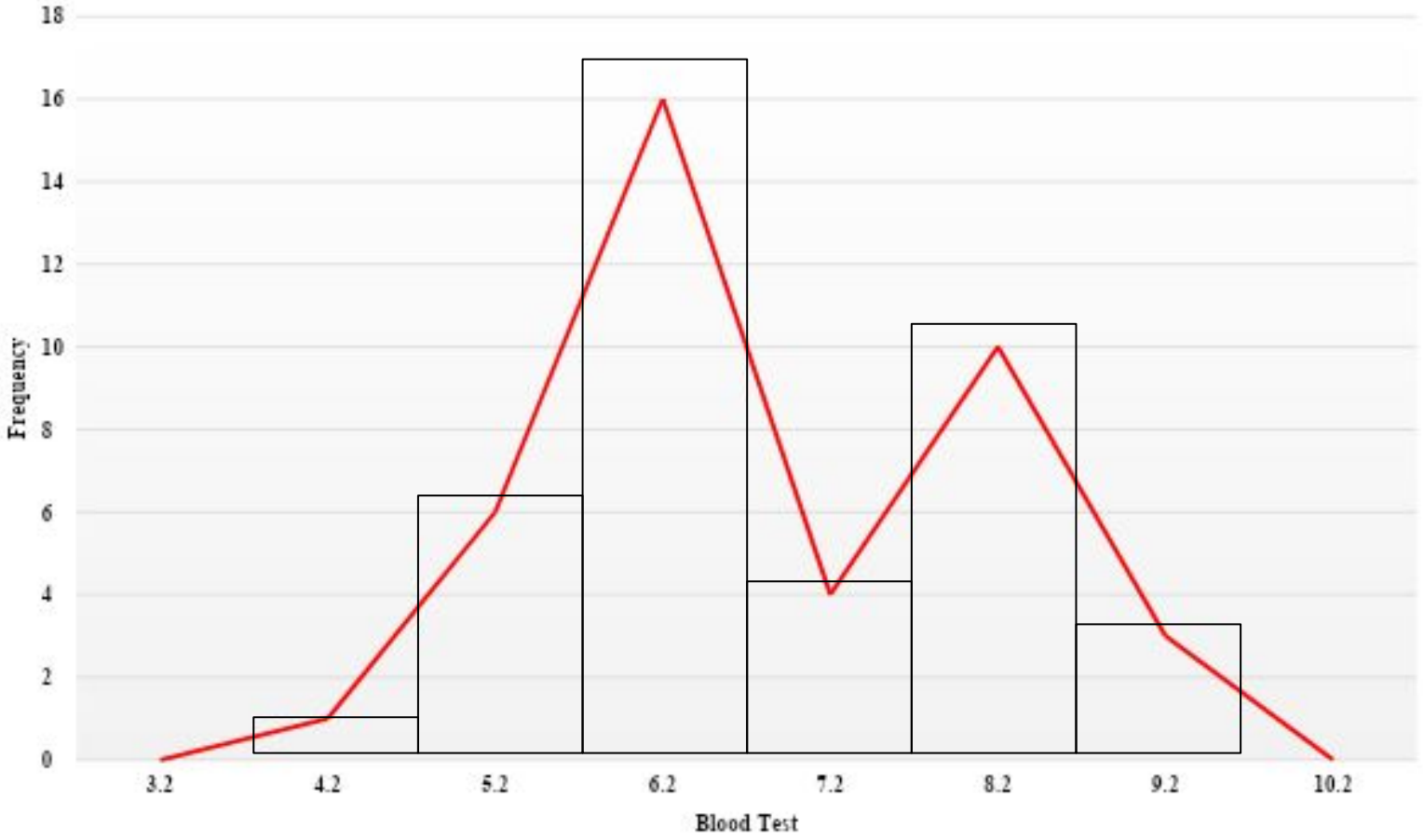
“just" right.

Bar Chart VS. Histogram

- With bar charts, each column represents a group defined by a categorical variable. With histograms, each column represents a group defined by a quantitative variable.
- With bar charts, however, the X axis does not have a low end or a high end; because the labels on the X axis are categorical - not quantitative. As a result, it is not appropriate to comment on the skewness of a bar chart.

Frequency Polygon

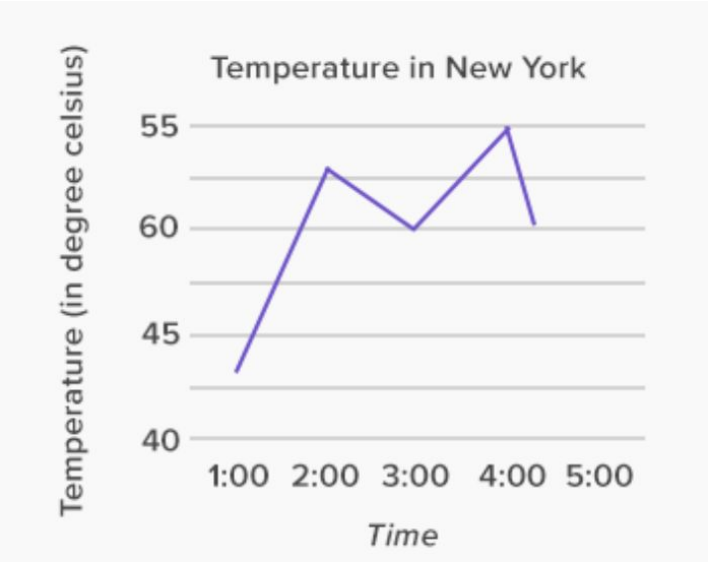
- A frequency polygon is a line graph of class frequency plotted against class mid-point.



Class Boundaries	Frequency f	Relative Frequency	Cumulative Rel. Frequency	Class Midpoint, x
3.7 - 4.7	1	0.025	0.025	4.2
4.7 - 5.7	6	0.150	0.175	5.2
5.7 - 6.7	16	0.400	0.575	6.2
6.7 - 7.7	4	0.100	0.675	7.2
7.7 - 8.7	10	0.250	0.925	8.2
8.7 - 9.7	3	0.075	1.000	9.2

Line Graph

- Line diagrams are used to show the trend of events with the passage of time.

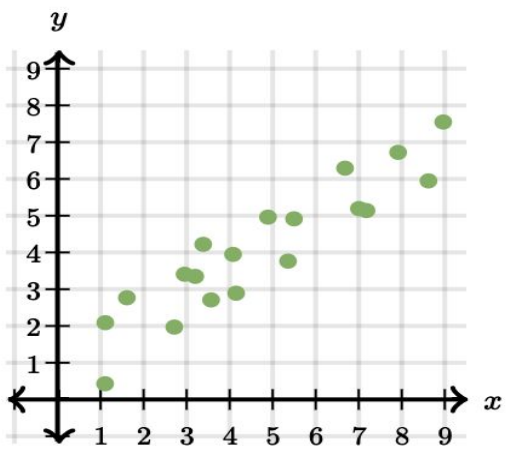


Scatterplot

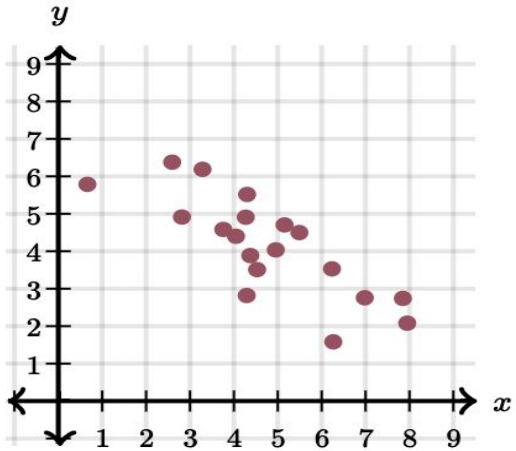
Scatterplot: A type of data display that shows the relationship between two numerical variables.

- The position of each dot on the horizontal and vertical axis indicates values for an individual data point
- Used to observe relationships between variables.

Positive correlation



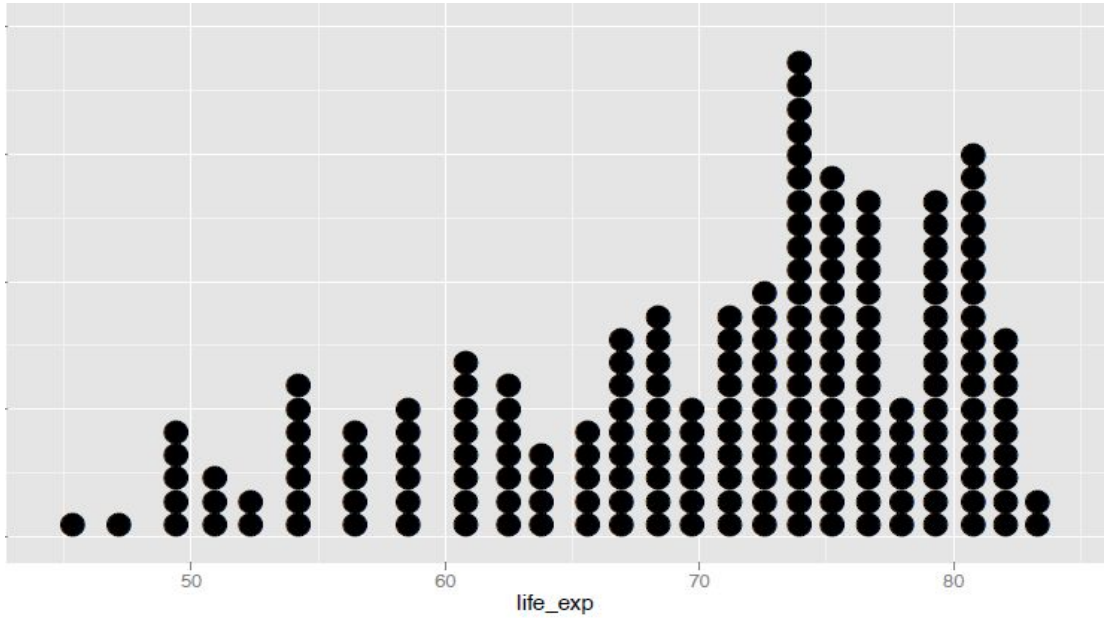
Negative correlation



Dotplot

Dot plot

- is a one variable scatter plot.
- useful if you want to investigate each variable separately.



stacked version

Thank You