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.

Types of Variables

Variables can be classified as

- **Qualitative (or Categorical)**
- **Quantitative (or Numerical)**

- Qualitative variables take on values that are names or labels. Numerical measurement are not possible. The color of a ball (e.g., red, green, blue) or the breed of a dog (e.g., collie, shepherd, terrier) would be examples of qualitative or categorical variables.

- Quantitative variables are numeric. They represent a measurable quantity. For example, when we speak of the population of a city, we are talking about the number of people in the city - a measurable attribute of the city. Therefore, population would be a quantitative variable.

Level of Measurement

Level of measurement defines the amount of information contained in the data. According to measurement scale:

- Nominal
- Ordinal
- Interval Scales
- Ratio Scales

Nominal Variables

- Nominal variable – a categorical variable *without* an intrinsic (general) order
- Examples of nominal variables:
  - Where a person lives in the U.S. (Northeast, South, Midwest, etc.)
  - Gender (Male, Female)
  - Nationality (American, Mexican, French)
  - Race/ethnicity (African American, Hispanic, White, Asian American)

Ordinal Variables

- Ordinal variable—a categorical variable with some intrinsic order

- Examples of ordinal variables:
  - Agreement (strongly disagree, disagree, neutral, agree, strongly agree)
  - Rating (excellent, good, fair, poor)
  - Frequency (always, often, sometimes, never)

Interval Scales

- 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.

Ratio Scales

- 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.

Numerical Variables (Cont..)

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

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).

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.

Introduction

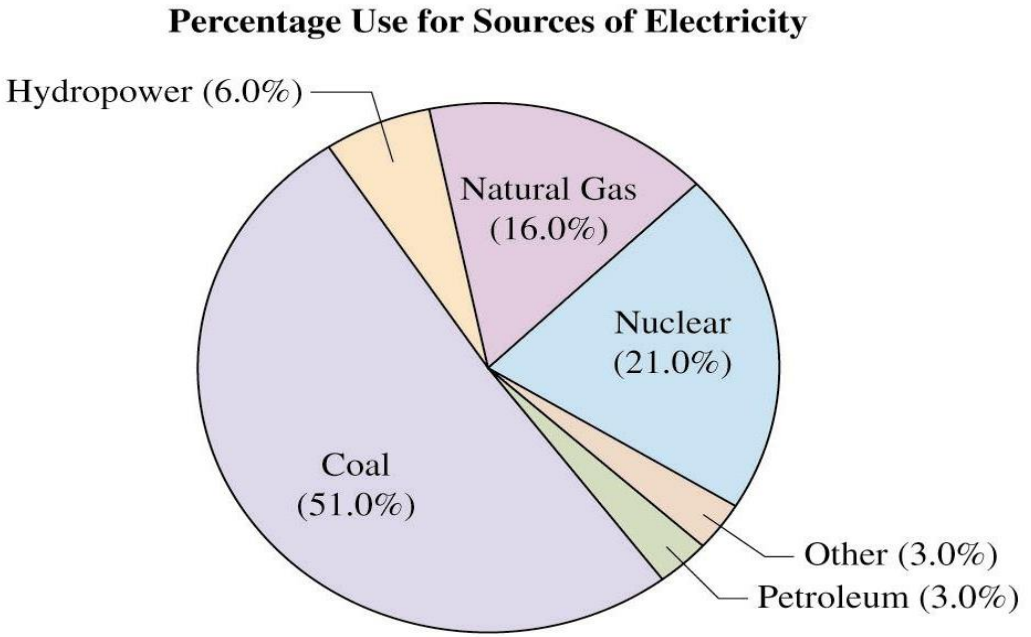
- An effective presentation of the data often quickly reveals important features such as
  - their range,
  - degree of symmetry,
  - how concentrated or spread out they are,
  - where they are concentrated, and so on.

- In this lecture we will be concerned with techniques, both tabular and graphic, for presenting data sets.
  - Graph for Categorical Variables
  - Frequency tables
  - The histogram
  - The stem-and-leaf plot
  - The scatter diagram

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

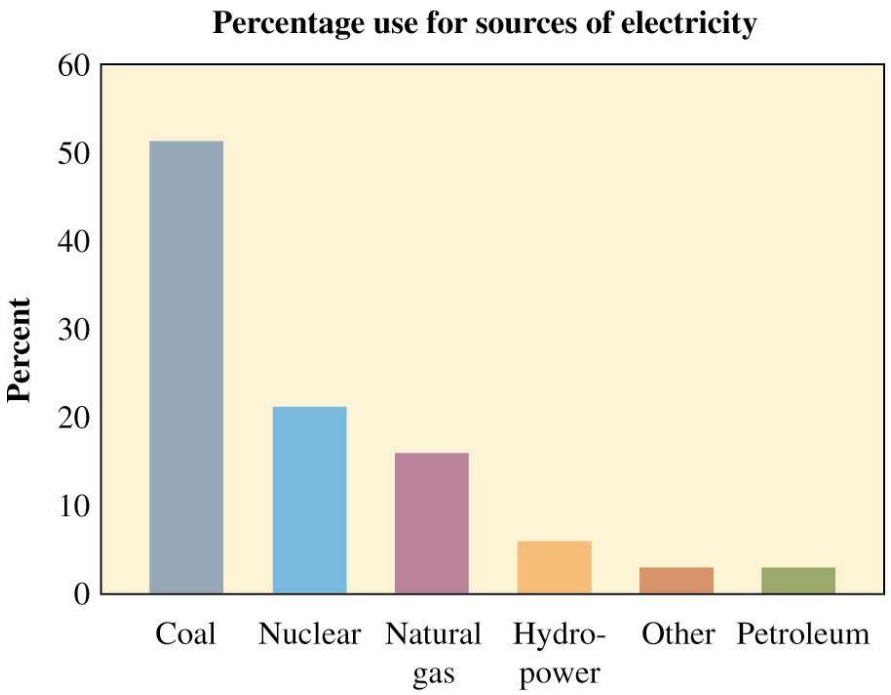
Source	U.S. Percentage
Coal	51
Hydropower	6
Natural gas	16
Nuclear	21
Petroleum	3
Other	3
Total	**100**



Bar Graphs

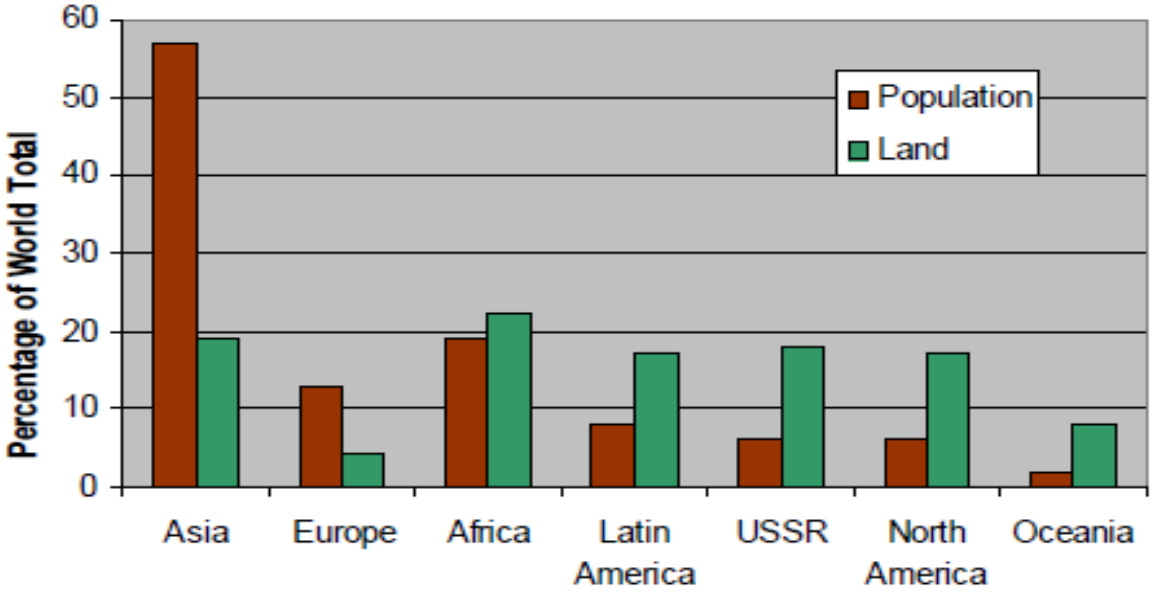
- Summarizes categorical variable
- Vertical bars for each category
- Height of each bar represents either counts or percentages
- Easier to compare categories with bar graph than with pie chart
- Called Pareto Charts when ordered from tallest to shortest

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



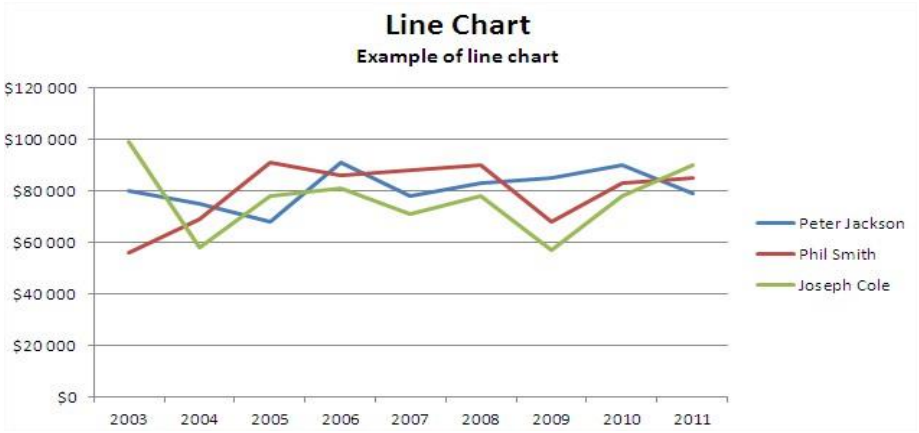
Multiple Bar Graphs

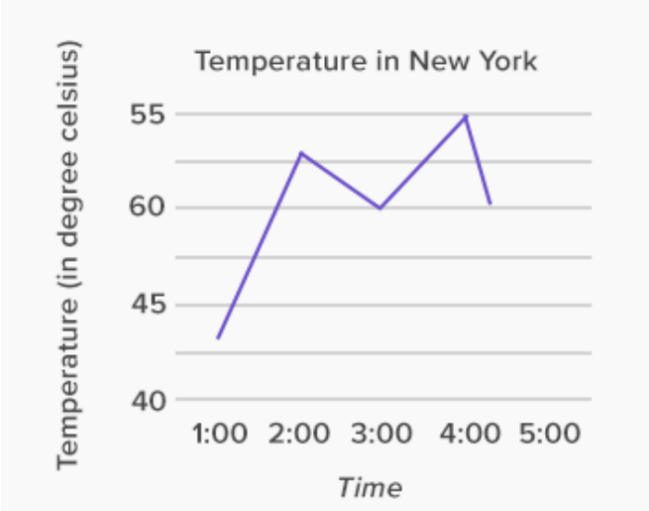
- Also called compound bar charts
- More then one sub-attribute of variable can be expressed
- Used for compare data



Line Graph

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





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, $4 + 5 + 8 = 17$.

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

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

Grouped Frequency Distribution

- Group the values into a set of classes
- A table that summarizes data by classes, or class intervals
- In a typical grouped frequency distribution, there are usually 5-12 classes of equal width
- The table may contain columns for class number, class interval, frequency, relative frequency, cumulative relative frequency, and class midpoint

Grouped Frequency Distribution (Cont…)

Guidelines for constructing a frequency distribution:

- For some data sets the number of distinct values is too large to utilize.
- In such cases, we divide the values into groupings, or class intervals.
- The number of class intervals chosen should be a trade-off between

  (1) choosing too few classes at a cost of losing too much information about the actual data values in a class and

  (2) choosing too many classes, which will result in the frequencies of each class being too small for a pattern to be discernible.
- Generally, 5 to 10 class intervals are typical.

Grouped Frequency Distribution (Cont…)

Guidelines for constructing a frequency distribution:

- All classes should be of the same width
- Classes should be set up so that they do not overlap and so that each piece of data belongs to exactly one class
- 5-12 classes are most desirable. The square root of *n* is a reasonable guideline for the number of classes if *n* is less than 150.

Grouped Frequency Distribution (Cont…)

- Class boundaries
  - the endpoints of a class interval
- 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

Table 2.5 Blood Cholesterol Levels

213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Table 2.7 Frequency Table of Blood Cholesterol Levels

Class intervals	Frequency	Relative frequency
170–180	3	$\frac{3}{40} = 0.075$
180–190	7	$\frac{7}{40} = 0.175$
190–200	13	$\frac{13}{40} = 0.325$
200–210	8	$\frac{8}{40} = 0.20$
210–220	5	$\frac{5}{40} = 0.125$
220–230	4	$\frac{4}{40} = 0.10$

Table 2.6 Blood Cholesterol Levels in Increasing Order

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195,
196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Grouped Frequency Distribution (Cont…)

Procedure for constructing a frequency distribution:

1. Identify the high (*H*) and low (*L*) scores. Find the range. Range = *H* - *L*

2. Select a number of classes and a class width so that the product is a bit larger than the range

3. Pick a starting point a little smaller than *L*. Count from *L* by the width to obtain the class boundaries. Observations that fall on class boundaries are placed into the class interval to the right.

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

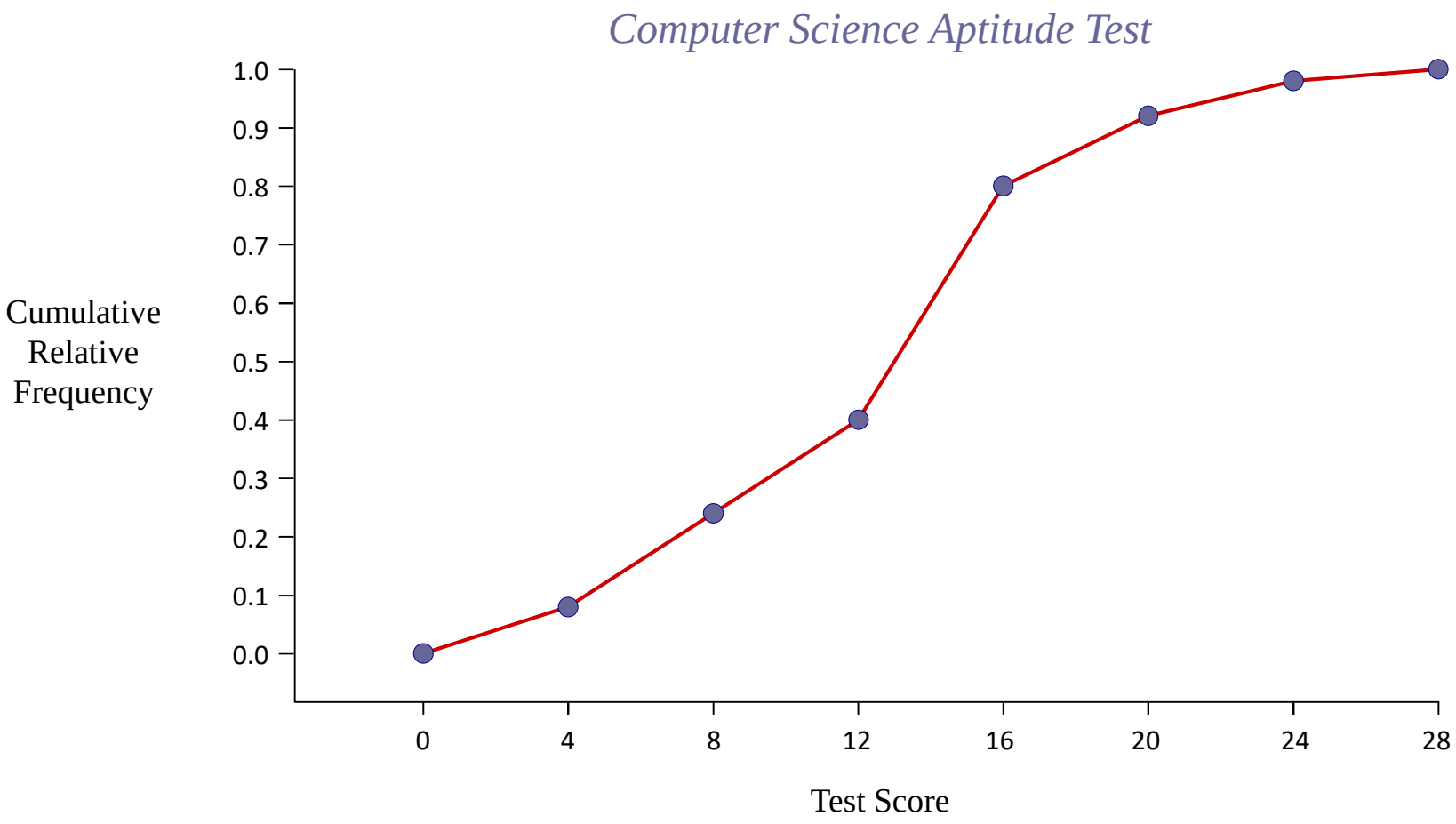
Ogive: A line graph of a cumulative frequency or cumulative relative frequency distribution. An ogive has the following components:

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:



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

Note: The display could be constructed so that only five possible values (instead of ten) could fall in each stem. What would the stems look like? Would there be a difference in appearance?

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.

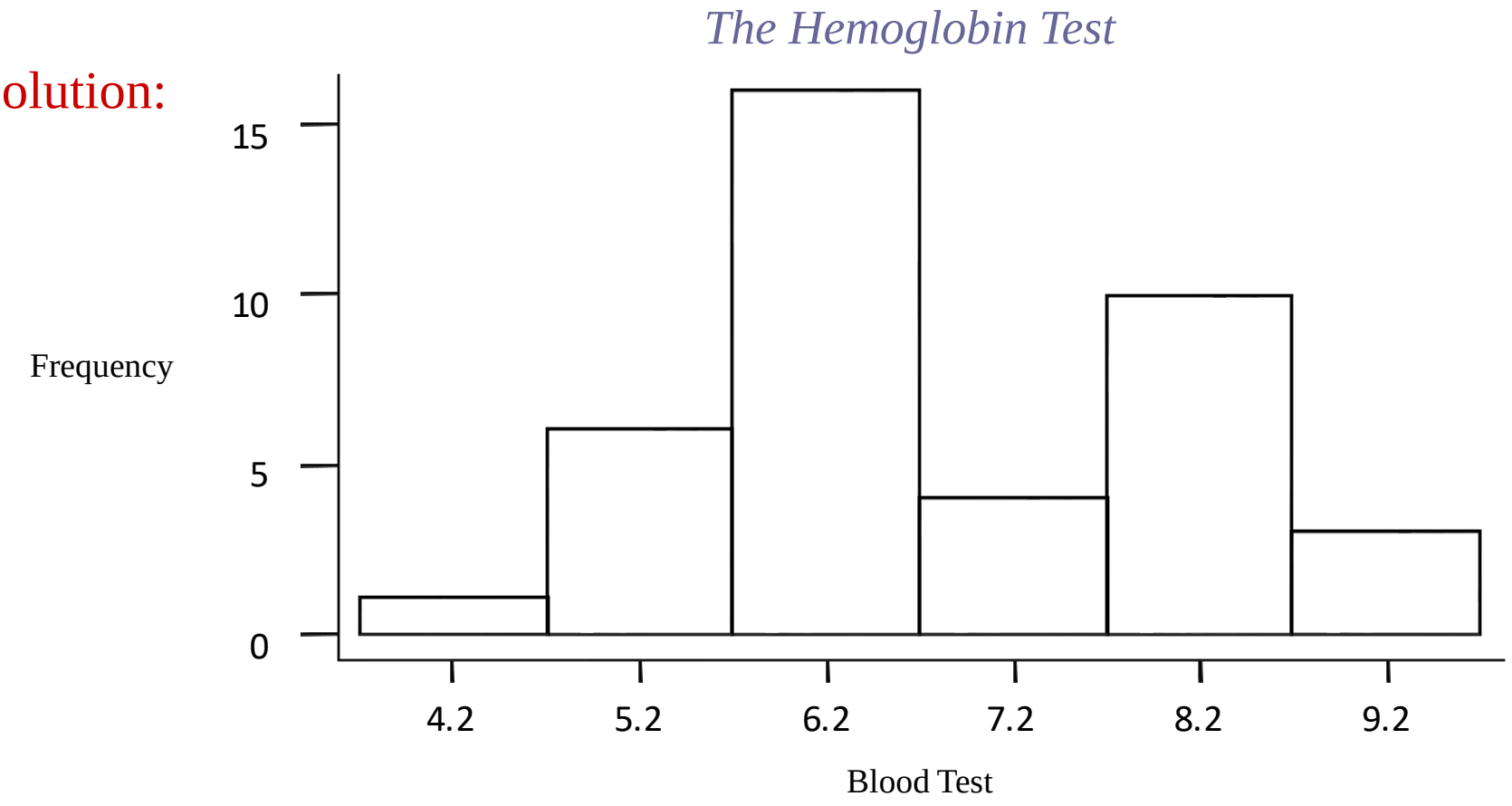
Notes:

- 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:

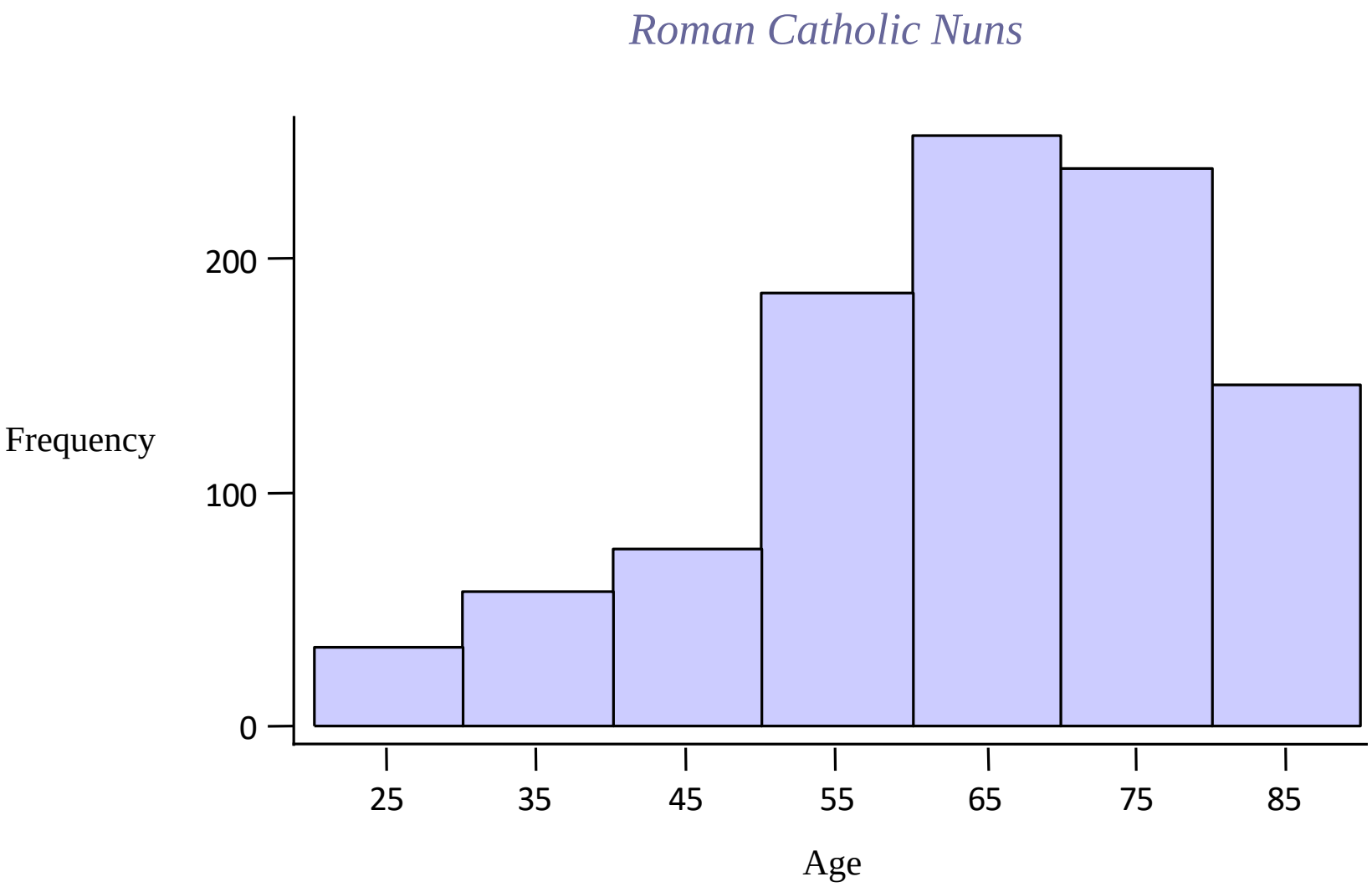


Example

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

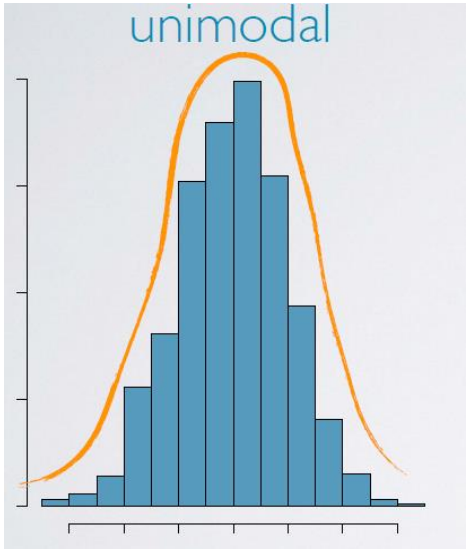
Age	Frequency	Class Midpoint
20 up to 30	34	25
30 up to 40	58	35
40 up to 50	76	45
50 up to 60	187	55
60 up to 70	254	65
70 up to 80	241	75
80 up to 90	147	85

Solution

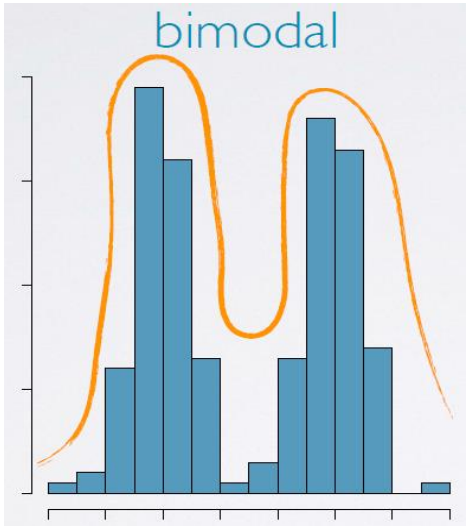

Roman Catholic Nuns
Frequency
200
100
0
25
35
45
55
65
75
85
Age

Modality of Histogram

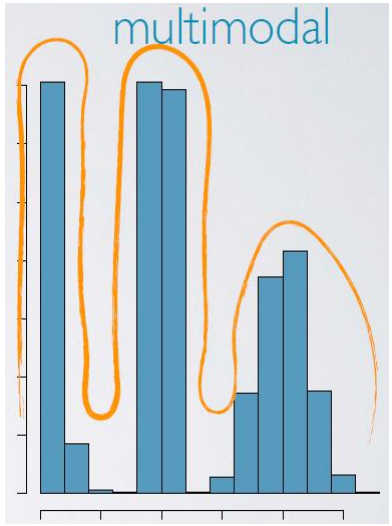
- Modality is also related to the shape of a histogram
- Refers to the number of prominent peaks



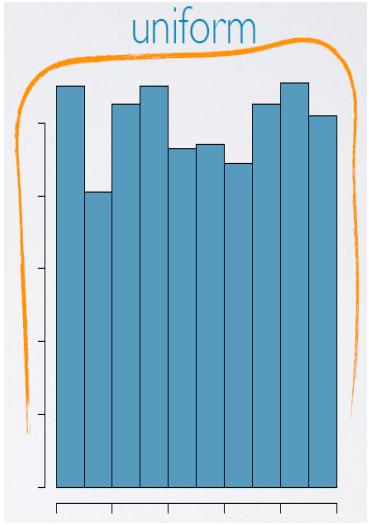
Uni-modal
one prominent peak



Bi-modal
two prominent peaks



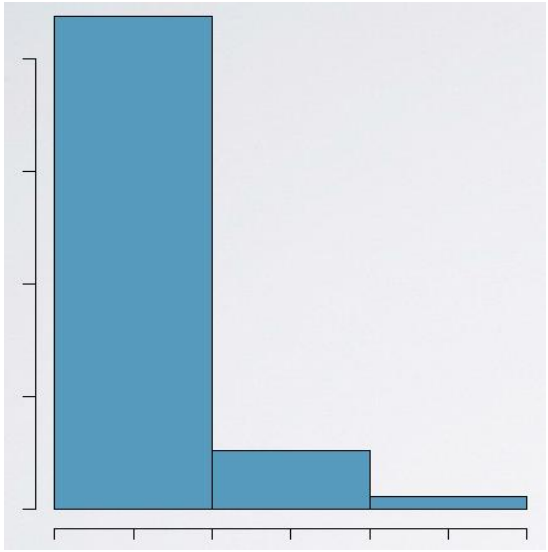
Multimodal
More than two prominent peaks



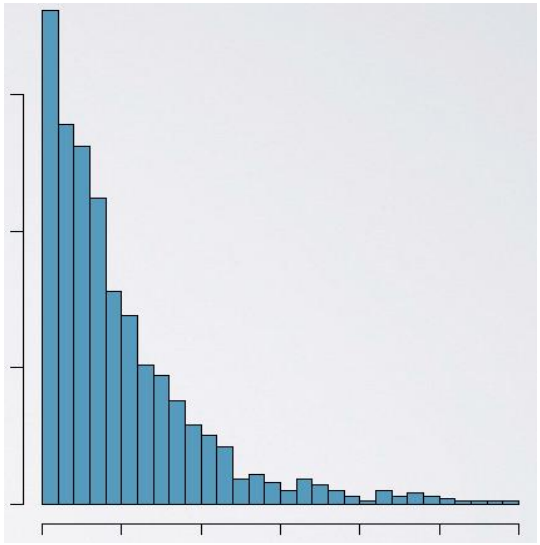
Uniform
no prominent peak

Bin Width of Histogram

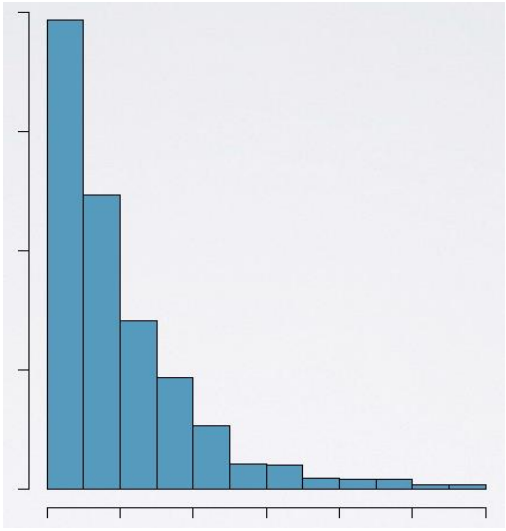
- The chosen bin width can alter the story that the histogram is telling.
- Too large bin width → we may lose interesting details.
- Too narrow bin width → difficult to get the overall picture of the distribution.
- Ideal bin width depends on the data being analyzed



too wide



too narrow



“just" right.

The Difference Between Bar Charts and Histograms

- With bar charts, each column represents a group defined by a categorical variable

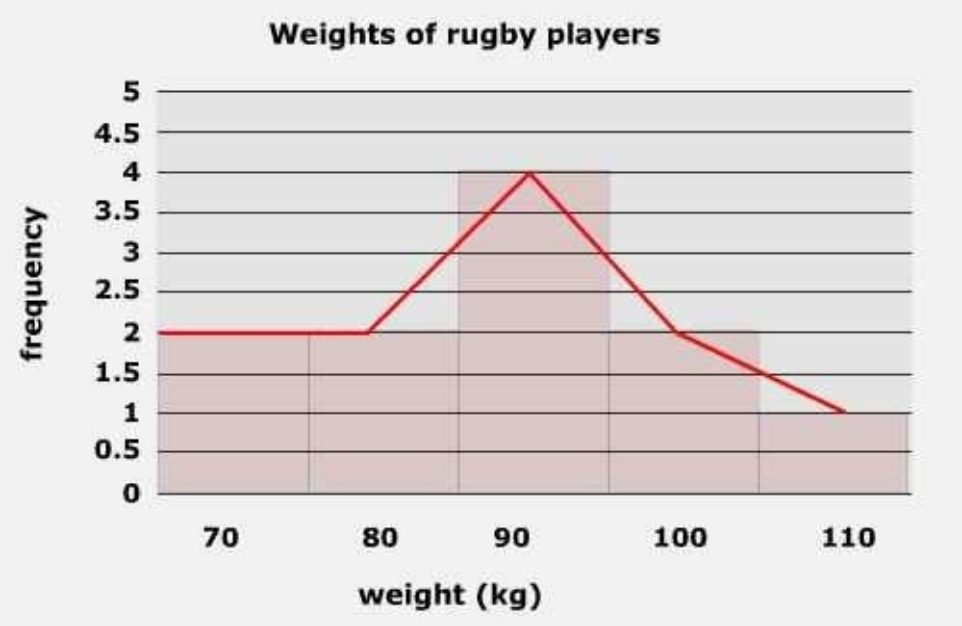
- With histograms, each column represents a group defined by a quantitative variable.

  - 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.

It can be obtained by connecting the mid-points of the tops of the rectangles in the histogram or drawn as a line graph.

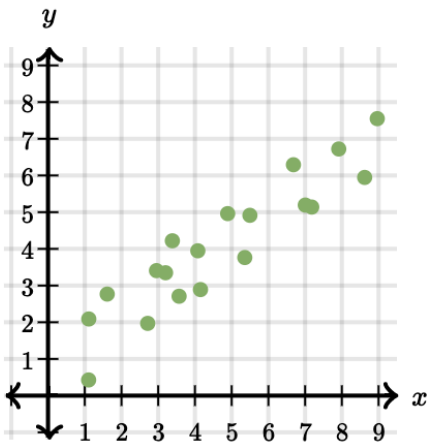


Scatterplot

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