Chapter 2

Section 2.1 What are the types of Data

Variable

A **variable** is any characteristic that is observed for the subjects in a study.

The data values that are observed for a variable are referred to as the **observations.**

Categorical and Quantitative variables

- **Categorical:** A variable is categorical if each observation belongs to one of a set of categories
- **Quantitative:** A variable is quantitative if observations on it take numerical values that represent different magnitudes of the variable.

Ex 7:

Examples for categorical variables

Gender: male, female

Hair color: blond, brown, red, black

Blood type: A, B, AB, O

Examples for quantitative variables

Age

Height

Number of children for a family

Discrete and Continuous variables

A quantitative variable could be either **discrete** or **continuous**.

- **Discrete:** A quantitative variable is discrete if it's possible values form a set of separate numbers such as 0,1,2,3,….
- **Continuous:** A quantitative variable is continuous if its possible values form an interval

Ex 8:

Examples for discrete variables

Number of students in a class: Only counting numbers

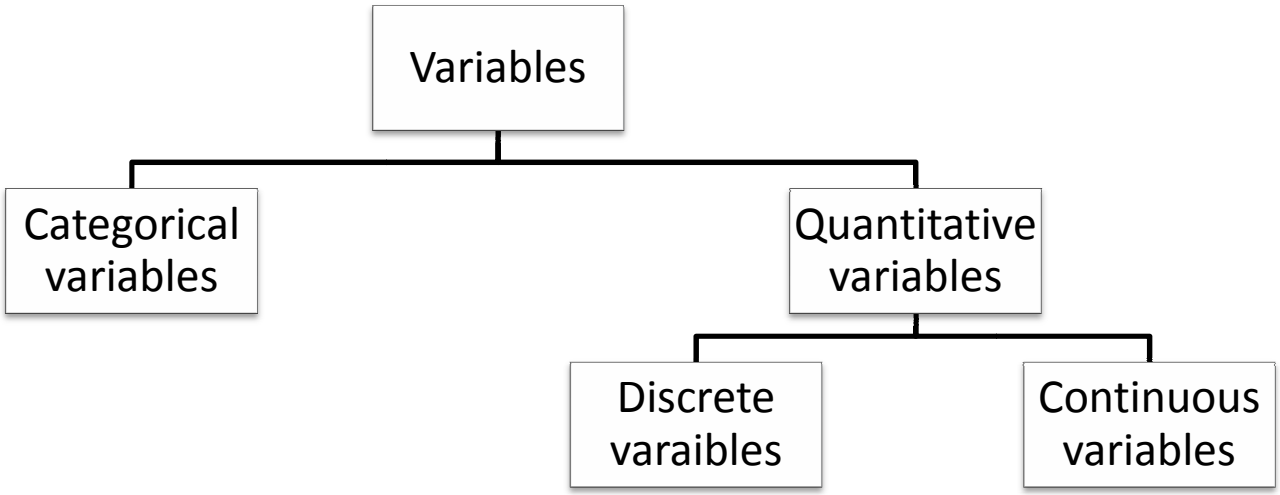
Height of a pile of bricks of height ½" each: Only multiples of ½"

Examples for continuous variables

Height of a person: Can take any value in a some interval (say from 1' to 9')

Temperature of Lubbock: Can take any value in some interval (25 to 110)

Classification of variables



Summarizing data using tables and graph

The methods used for summarizing and analyzing data depend on the type of the variable.

Proportion and Percentages (relative frequencies)

- **Proportion** of the observations that fall in a certain category
  $= \frac{frequency\ (count)\ of\ observations\ in\ that\ category}{total\ number\ of\ observations}$
- **Percentage** of the observations that fall in a certain category = **Proportion * 100**
  $= \frac{frequency\ (count)\ of\ observations\ in\ that\ category}{total\ number\ of\ observations} \times 100$

Proportions and percentages are also called **relative frequencies**.

Frequency Table

A **frequency table** is a listing of possible values for a variable, together with the number of observations for each value.

Ex 9:

A campus press polled a sample of 300 undergraduate students in order to study students' attitude towards a proposed change in a dormitory regulation. Summary of results of an opinion poll is as follows.

Response	Frequency	Proportion	Percentage
Support	150	=150/300 = 0.5	50%
Neutral	50	= 50/300 = 0.167	16.7%
Oppose	100	= 100/300 = 0.333	33.3%
Total	**300**	**1**	**100%**

A survey was done in a small town about the number of stories of building. The following frequency table summarizes the findings.

Number of stories	Frequency	Proportion	Percentage
1	100	=100/200 = 0.5	50%
2	50	= 50/200 = 0.25	25%
3	50	= 50/200 = 0.25	25%
Total	**200**	**1**	**100%**

Section 2.2 Describing data using Graphical Summaries

Graphs for Categorical Variables

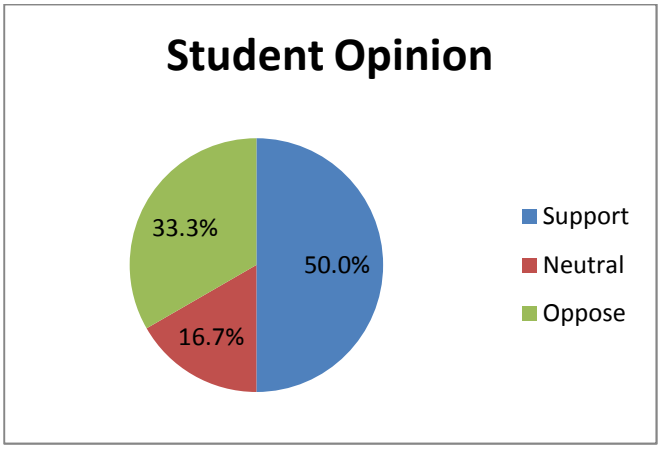
Two methods are often used to present categorical data graphically.

- Pie chart
- Bar graph

Pie chart

A **pie chart** is a circle having a "slice of the pie" for each category. The size of a slice corresponds to the percentage of observations in the category.

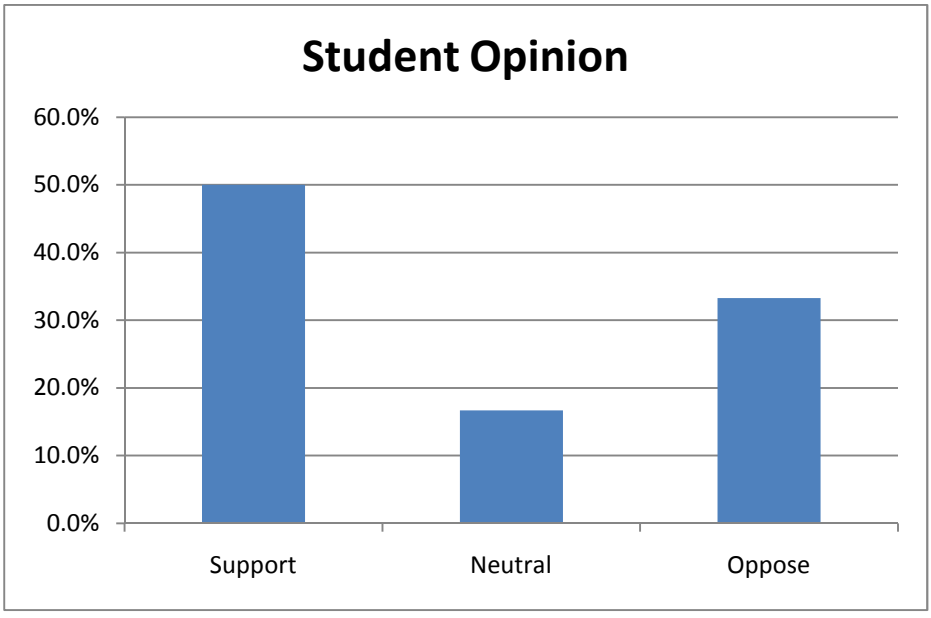
Ex 10: Referring back to Ex 9 in the previous section, the pie chart for the scenario looks like,



Bar graph

A **bar graph** displays a vertical bar for each category. The height of the bar is the percentage of observations in the category.

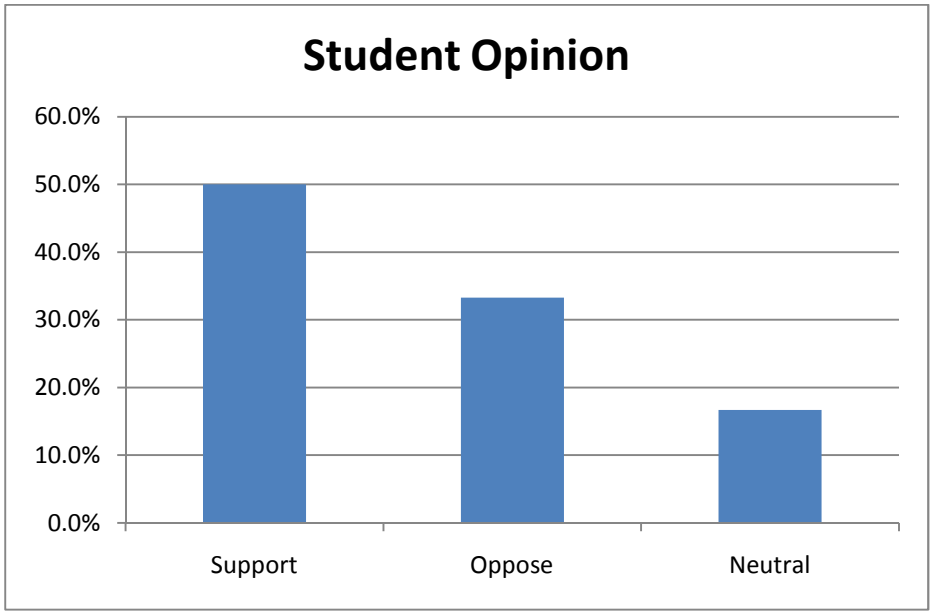
Ex 11: Referring back to Ex 9 again, the bar chart for the same scenario looks like,



Pareto Charts

A **Pareto chart** is the usual bar chart with the columns ordered by the frequency, from the tallest bar to the shortest.

Ex 12: The Pareto chart scenario under consideration looks like,



Graphs for Quantitative Variables

Following methods are used for graphically summarizing quantitative data.

- Dot plot
- Stem-and-Leaf plot
- Histogram

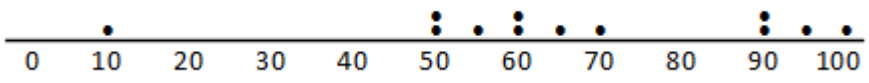
Dot plot

Dot plots are constructed by

- Drawing a horizontal line labeled with the variable name
- Marking regular values of the variable on it
- Placing a dot for each observation above the corresponding value

Ex 13: The following set of data is the scores obtained for midterm test on a 0-100 scale. Construct a dot plot.

10, 90, 95, 100, 65, 50, 60, 50, 90, 55, 60, 70



Advantages

- Can identify unusual data (outliers)
- Can identify concentrated points since the number of dots represent the frequency of occurrence of that value

Stem-and-Leaf plot

- Each observation is represented by a **stem** and a **leaf**. Usually the stem consists of all the digits except for the final one, which is the leaf.
- Place the stems in the left hand side starting from the smallest
- List the leaves on the right in increasing order

Ex 14: Math 1320 midterm exam scores of 20 students are given below. Create a stem-and-leaf plot for these scores.

80	85	75	90	62	50	55	65	75	82
70	25	92	57	63	72	81	95	41	69

Stem	Leaf
2	5
3	
4	1
5	057
6	2359
7	0255
8	0125
9	025

Histogram

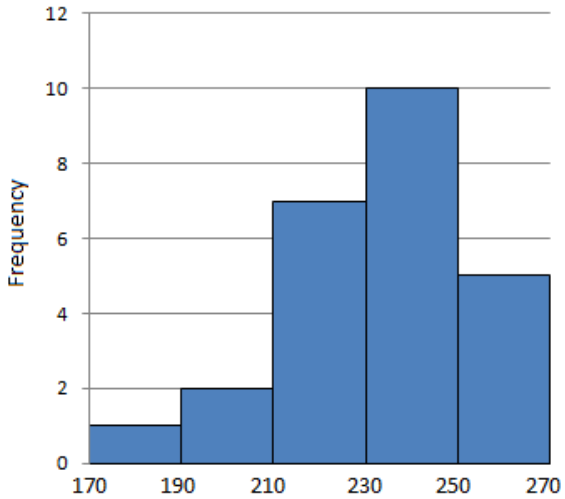
A **histogram** is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

Steps

- Divide the range of the data into intervals of equal width. (For discrete variables with few values, use the actual possible values.
- Count the number of observations (the frequency) in each interval, forming a frequency table.
- On the horizontal axis, label the values or the endpoints of the intervals. Draw a bar over each value or interval with height equal to its frequency (or percentage).

Ex 15: Construct a histogram for the data given below.

Interval	Frequency
170-189.9	1
190-209.9	2
210-229.9	7
230-249.9	10
250-269.9	5



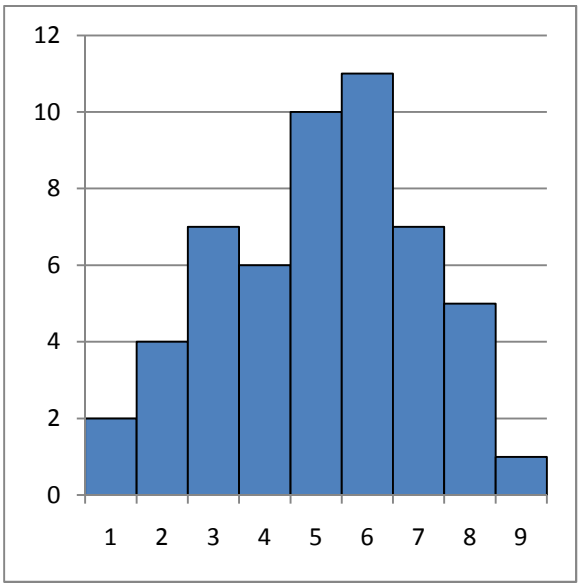
Note: The dot plots and stem-and-leaf plots are more useful with small data sets. (It is not feasible to construct dot plots or stem-and-leaf plots for large data sets since they show each and every observation).

In contrast, histograms can be used for larger data sets as it is more compact. (It does not display each and every observation).

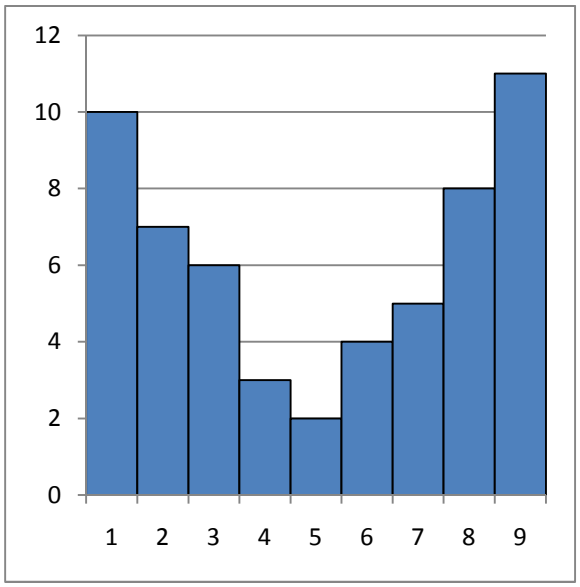
Shapes of Distributions

One of the purposes of making graphs is to identify the patterns of data distributions. The following distributions are sometimes common in practice.

Unimodal



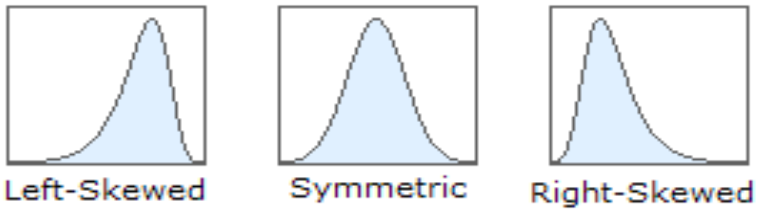
Bimodal

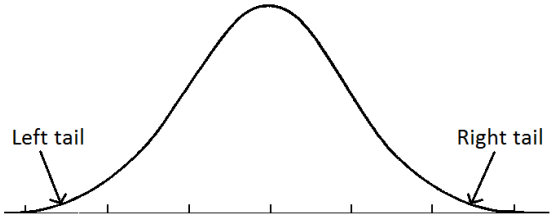


Skewed distributions

A distribution is

- **Skewed to the left** if the left tail is longer than the right tail
- **Skewed to the right** if the right tail is longer than the left tail





Numerical Summaries of Quantitative data

In the next two sections, we are going to consider two types of numerical summaries for quantitative data. They are,

- **Center**
- **Spread**

of the distribution.

Section 2.3 Describing the Center of Quantitative data

There are two commonly used measures for the center of a distribution.

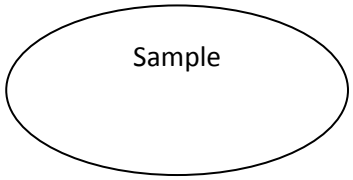
- **Mean**
- **Median**

Mean

The **mean** is the sum of the observations divided by the number of observations.

Notation

Population mean: μ



An observation: x

A set of observations: $x_1, x_2, x_3, \cdots, x_n$

Sample mean $(\bar{x}) = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

$$\bar{x} = \frac{\sum x}{n}$$

Median

The **median** is the midpoint of the observations when they are ordered from the smallest to the largest (or from the largest to the smallest).

If the number of observations is odd

There is a specific middle observation in the ordered set. This middle observation is the median

12 14 15 17 (20) 24 24 27 29

Median

If the number of observations is even

There are two middle observations. The median is the average of these two middle observations.

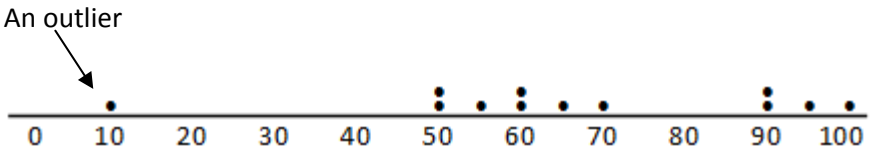
12 14 15 17 **20 24** 24 27 29 30

Median $= \frac{20+24}{2} = 22$

Outlier

An **outlier** is an observation that falls well above or well below the overall bulk of the data.

Ex 16:



Influence of an outlier

- The mean can be highly influenced by an outlier
- The median is not affected by an outlier

Ex 17:

Calculate the mean and the median of the following data set. 70, 46, 623, 64, 15

623 is an outlier for this data set.

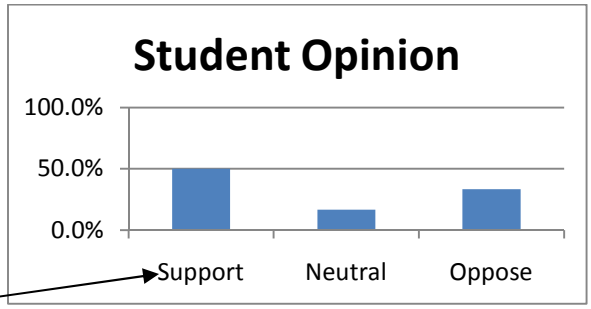
Mean $= \frac{70+46+623+64+15}{5} = 163.6$

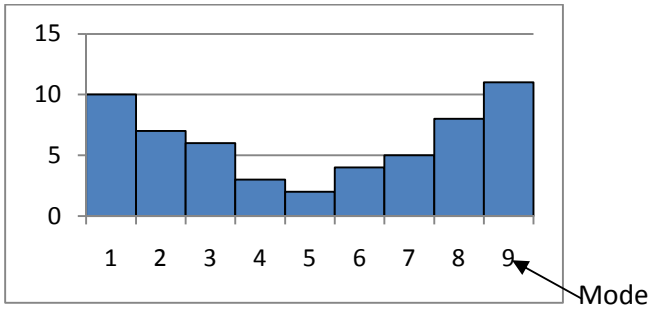
Median = 64

We can see from this example that the mean can be highly influenced by one outlier whereas the median is not affected by it. (In other words, the median lies close to the general trend whereas the mean is deviated out from the general trend).

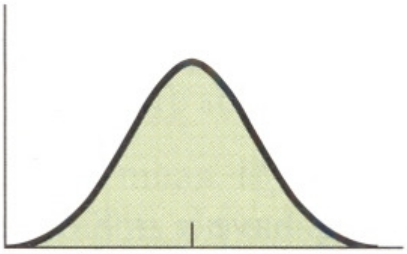
Mode

- The **mode** is the value that occurs most frequently. This need ***not*** necessarily be close to the center of the data set.
- Mode is most often used with categorical variables or discrete variables with small number of possible values.



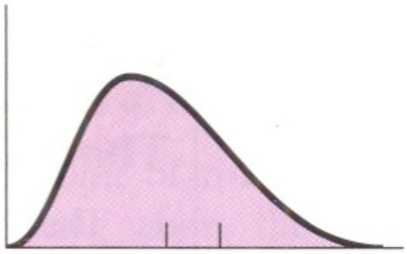


Symmetric Distribution



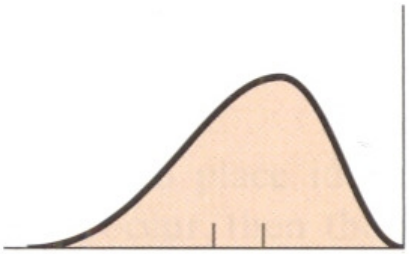
Mean = Median

Right-Skewed Distribution



Median Mean

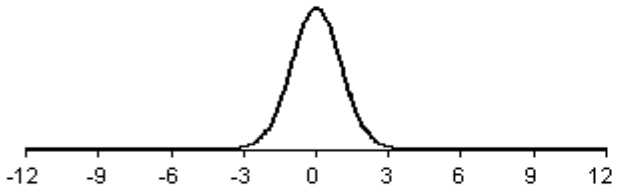
Left-Skewed Distribution



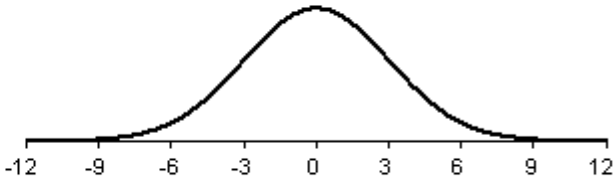
Mean Median

Section 2.4 Describing the Spread of Quantitative data

This measures how much the data set is spread.



Lower spread



Higer spread

There are two main measures of spread. Spread is about how much the data set is spread out. They are,

- **Range**
- **Standard Deviation**

Range

Range is the difference between the largest and the smallest observations.

Ex 18: Find the range of the following set of data. 70, 46, 623, 64, 15

Range = 623 – 15 = 608

Standard Deviation

Deviation: The **deviation** of an observation x from its mean $\bar{x}$ is $(x - \bar{x})$; i.e. the difference between the observation and the sample mean.

Note: If the observation under consideration is larger than the mean, then the deviation is positive whereas if the observation is smaller than the mean, the deviation is negative.

Standard Deviation

The **standard deviation (s)** of n observations is

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}} = \sqrt{\frac{sum\ of\ squared\ deviations}{sample\ size - 1}}$$

$\boldsymbol{s^2}$ is called the **variance**. $\left(s^2 = \frac{\sum(x-\bar{x})^2}{n-1}\right)$

Ex 19: Find the standard deviation and the variance for the following set of data. 4, 5, 6, 7, 8

$$\bar{x} = \frac{\sum x}{n} = \frac{4 + 5 + 6 + 7 + 8}{5} = 6$$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1} = \frac{(4-6)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 + (9-6)^2}{5 - 1} = \frac{4 + 1 + 0 + 1 + 4}{4} = \frac{10}{4} = 2.5$$

Variance

Standard deviation $s = \sqrt{2.5} = 1.581$

Alternate method

x	$(x - \bar{x})$	$(x - \bar{x})^2$
4	-2	4
5	-1	1
6	0	0
7	1	1
8	2	4
$\sum x = 30$		$\sum (x - \bar{x})^2 = 10$

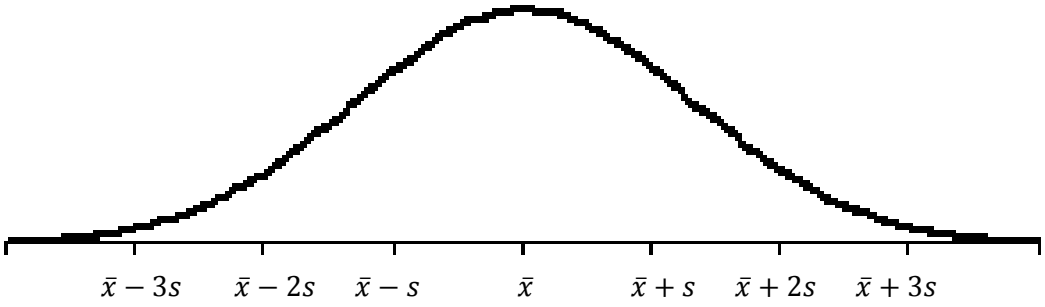
$$\bar{x} = \frac{\sum x}{n} = \frac{30}{5} = 6$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} = \frac{10}{4} = 2.5$$

$$s = \sqrt{2.5} = 1.581$$

Empirical Rule

- 68% of the data fall within 1 standard deviation from the mean
- 95% of the data fall within 2 standard deviation from the mean
- Almost all of the data fall within 3 standard deviation from the mean



Notation

Population

Standard deviation: σ

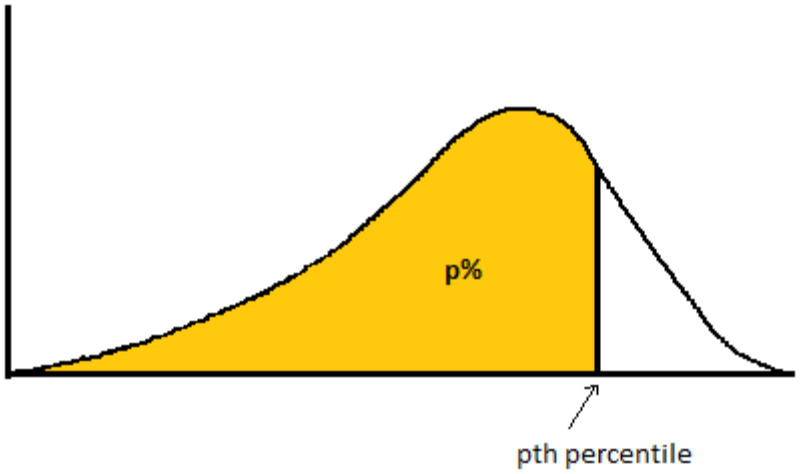
Sample

Sample standard deviation: s

Section 2.5 Describing the spread with Measure of Position

Percentile

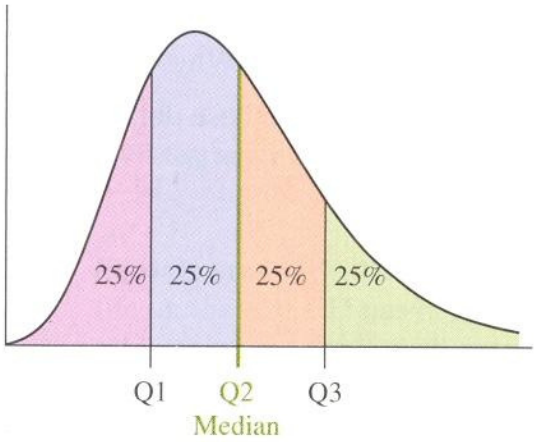
The ***pth* percentile** is a value such that *p* percent of the observations fall below or at that value.



Ex 20: SAT score 1000 and percentile = 70% means that 70% of the test takers have a score less than or equal to 1000.

Quartiles

- **Q1 (First quartile)** is the 25th percentile
- **Q2 (Second quartile)** is the 50th percentile = Median
- **Q3 (Third quartile)** is the 75th percentile



Finding the quartiles

- Order the dataset
- Find Q2 which is the median
- Then Q1 is the median of the data set to the left of Q2
- Q3 is the median of the data set to the right of Q2

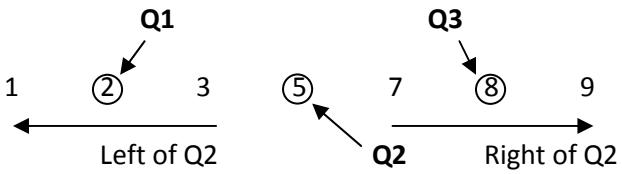
Note: If the number of observations in the dataset is odd, the middle observation is not considered in step 3 and 4.

Ex 21:

Odd # of observations: Consider the following dataset: 1 7 2 9 8 3 5

Find the quartiles

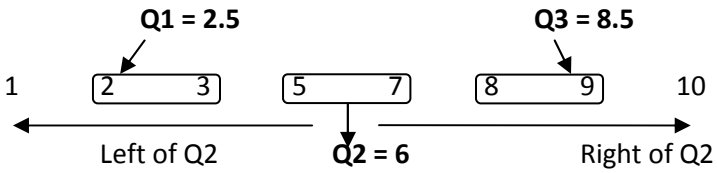
Ordered dataset:



Even # of observations: Consider the following dataset: 1 7 2 9 8 3 5 10

Find the quartiles

Ordered dataset:



Interquartile Range (IQR)

The **interquartile range** is the distance between the third and the first quartiles.

$$IQR = Q3 - Q1$$

Ex 22:

IQR for **odd** case in Ex 21 $= 8 - 2 = 6$

IQR for **even** case in Ex 21 $= 8.5 - 2.5 = 6$

The 1.5 * IQR criterion for Identifying Potential Outliers

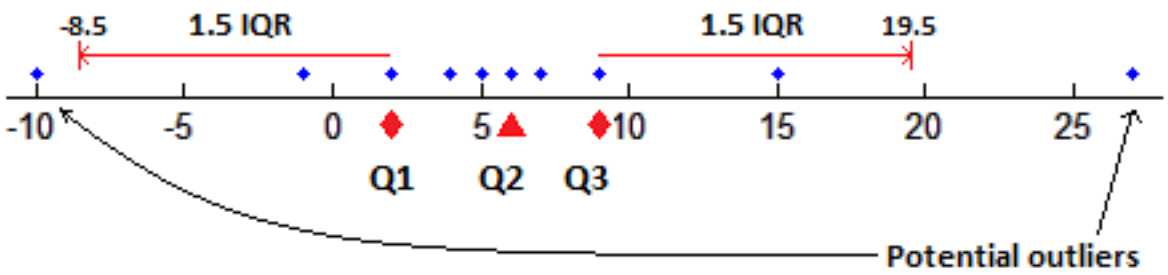
An observations is a **potential outlier** if it falls more than 1.5 * IQR below the first quartile or more than 1.5 * IQR above the third quartile.

Ex 23: Find the potential outliers for the dataset

-10 -1 2 2 4 5 6 7 9 15 27

Median (Q2) = 5; Q1 = 2; Q3 = 9; IQR = 9 – 2 =7; 1.5 * IQR = 10.5

Lower limit $= Q1 - 1.5\,IQR = 2 - 10.5 = -8.5$ Upper limit $= Q3 + 1.5\,IQR = 9 + 10.5 = 19.5$



Five Number Summary of Position

Five number summary of a dataset is

- Minimum value
- First quartile
- Median
- Third quartile
- Maximum value

Box Plot

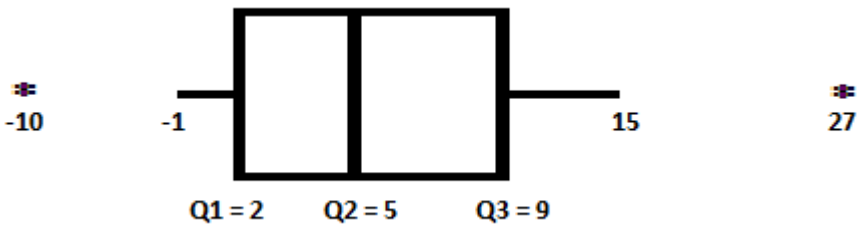
The **box plot** is a graphical display of the five number summary.

Construction

- Draw a box from Q1 to Q3
- A line is drawn inside the box to show Q2
- Two lines from the two sides of the box are extended until the smallest and the largest data points which are not potential outliers
- Potential outliers are shown separately

Ex 24: Draw a box plot for the data set in Ex 23.

{Median (Q2) = 5; Q1 = 2; Q3 = 9; IQR = 9 – 2 =7; 1.5 * IQR = 10.5}



Z-Score

The **z-score** for an observation is the number of standard deviations that it falls from the mean. i.e.,

$$z = \frac{observation - mean}{standard\ deviation} = \frac{x - \mu}{\sigma}$$

Ex 25: The mid-term exam score of a student is 90. The class average is 85 and the standard deviation is 5. Find the z-score.

$z = \frac{90-85}{5} = 1$