Statistics for Scientists – CSC261

Measurement of Central Tendency

Dr DEBAJYOTI PAL

SCHOOL OF INFORMATION TECHNOLOGY,
KMUTT

Summary Definitions

- The **central tendency** is the extent to which all the **data values group around a typical or central value**.

- The **variation** is the amount of **dispersion**, or **scattering**, of values

- The **shape** is the pattern of the distribution of values from the **lowest value to the highest** value.

Measures of Central Tendency: The Mean

- The arithmetic mean (often just called "mean") is the most common measure of central tendency
- For a sample of size n:

Pronounced x-bar

The i^{th} value

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Sample size

Observed values

Numerical Measures of Central Tendency

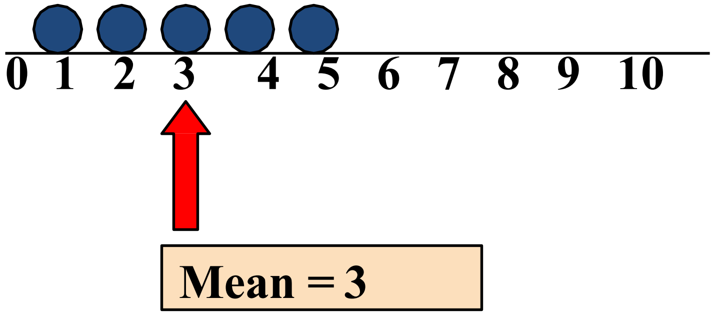
- The Mean
- Arithmetic average of the elements of the data set
- Sample mean denoted by x-bar, $\bar{x}$
- Population mean denoted by μ
- Calculated as

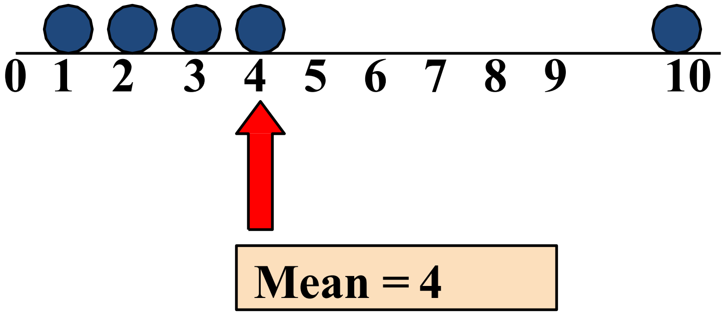
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \mu = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measures of Central Tendency: The Mean

(continued)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)





$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Mean for Grouped Data

Formula for Mean is given by $\bar{X} = \dfrac{\sum f(X)}{n}$

Where

$\bar{X}$ = Mean

$\sum f(X)$ = Sum of cross products of frequency in each class with midpoint X of each class

n = Total number of observations (Total frequency) = $\sum f$

Mean for Grouped Data Example

Find the arithmetic mean for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Solution for the Example

	A	B	C	D
1	Class	X (mid pt)	f	fX
2	0-1	0.5	1	0.5
3	1-2	1.5	4	6.0
4	2-3	2.5	8	20.0
5	3-4	3.5	7	24.5
6	4-5	4.5	3	13.5
7	5-6	5.5	2	11.0
8	Totals		25	75.5
9	Mean			3.02

Applying the formula $\bar{X} = \frac{\sum f(X)}{n}$ $= 75.5/25 = 3.02$

Mean

Class interval		f
0	49.99	78
50	99.99	123
100	149.99	187
150	199.99	82
200	249.99	51
250	299.99	47
300	349.99	13
350	399.99	9
400	449.99	6
450	499.99	4
		600

Weighted mean: A weighted mean is a kind of average. Instead of each data point contributing equally to the final mean, some data points contribute more “weight” than others.

To calculate an average that takes into account the importance of each value to the overall cost . Find out **average cost of labor per hour for each of the products.**

		Labor hrs per unit of output	
Grade of labor	Hourly wage	Product 1	Product 2
Unskilled	5	1	4
Semiskilled	7	2	3
Skilled	9	5	3

Weighted mean: A weighted mean is a kind of average. Instead of each data point contributing equally to the final mean, some data points contribute more "weight" than others.

To calculate an average that takes into account the importance of each value to the overall cost . Find out **average cost of labor per hour f**or each of the product

		Labor hrs per unit of output	
Grade of labor	Hourly wage	Product 1	Product 2
Unskilled	5	1	4
Semiskilled	7	2	3
Skilled	9	5	3

A simple arithmetic mean = (5+7+9) / 3= 7/hr
Using this, labor cost of 1 unit of product 1 to be = 7* (1+2+5) = 56
2 = 7* (4+3+3) = 70
Both are incorrect, the answers must take into account **that different amount of each grade of labor** .

P1= avg cost of labor per hr = (5*1+7*2+9*5)/8 = 8
P2= avg cost of labor per hr = (5*4+7*3+9*3)/10 = 6.80

Geometric Mean

- X= {1.09, 1.11, 0.98, 1.05, 1.01} = $(1.2574)^{0.2}$ =1.0468

Disadvantages of Mean:

It may be affected by extreme values

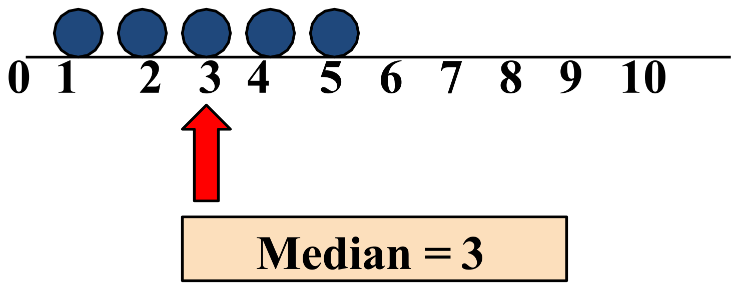
Tedious to compute

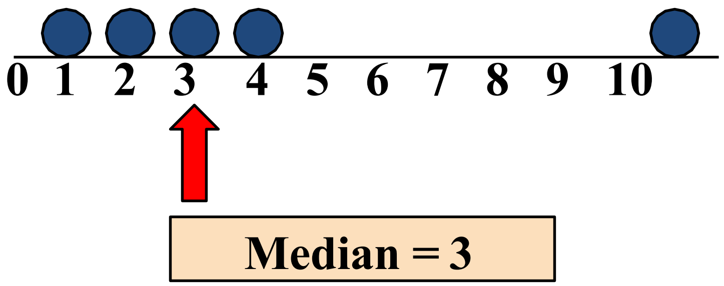
Cannot compute in case of open class

Cannot compute in case of categorical data

Measures of Central Tendency: The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)



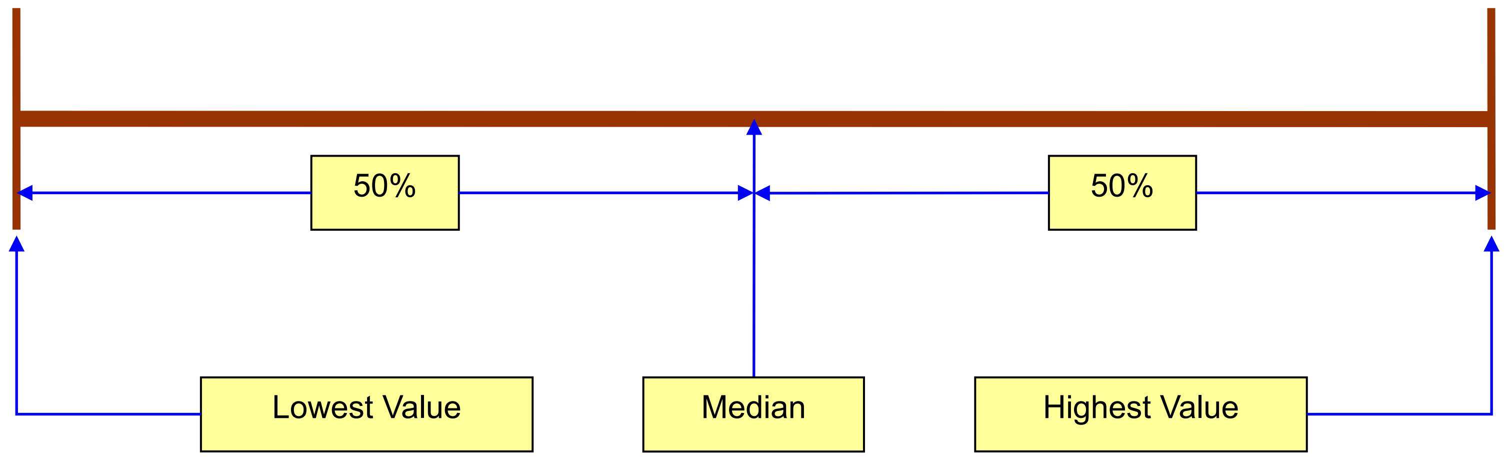


- Not affected by extreme values

Numerical Measures of Central Tendency

- The Median
- Middle number when observations are arranged in order, increasingly or decreasingly
- Median denoted by *m*
- Identified as the $\frac{n}{2}+0.5$ observation if *n* is odd, and the mean of the $\frac{n}{2}$ and $\frac{n}{2}+1$ observations if *n* is even

Numerical Measures of Central Tendency



Median for Grouped Data

Formula for Median is given by

$$\text{Median} = \boxed{L + \frac{(n/2) - m}{f} \times c}$$

Where

L =Lower limit of the median class

n = Total number of observations = $\boxed{\sum f(x)}$

m = Cumulative frequency preceding the median class

f = Frequency of the median class

c = Class interval of the median class

Median for Grouped Data Example

Find the median for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Solution for the Example

Class	Frequency	Cumulative Frequency
0-1	1	1
1-2	4	5
2-3	8	13
3-4	7	20
4-5	3	23
5-6	2	25
Total	**25**	

L =Lower limit of the median class
n = Total number of observations
m = Cumulative frequency **preceding** the median class
f = Frequency of the median class
c = Class interval of the median class

Substituting in the formula the relevant values,

Median = $L + \dfrac{(n/2) - m}{f} \times c$ we have Median = $2 + \dfrac{(25/2) - 5}{8} \times 1$

$= 2.9375$

Find median

Class interval		f
0	49.99	78
50	99.99	123
100	149.99	187
150	199.99	82
200	249.99	51
250	299.99	47
300	349.99	13
350	399.99	9
400	449.99	6
450	499.99	4
		600

L =Lower limit of the median class
n = Total number of observations
m = Cumulative frequency preceding the median class
f = Frequency of the median class
c = Class interval of the median class

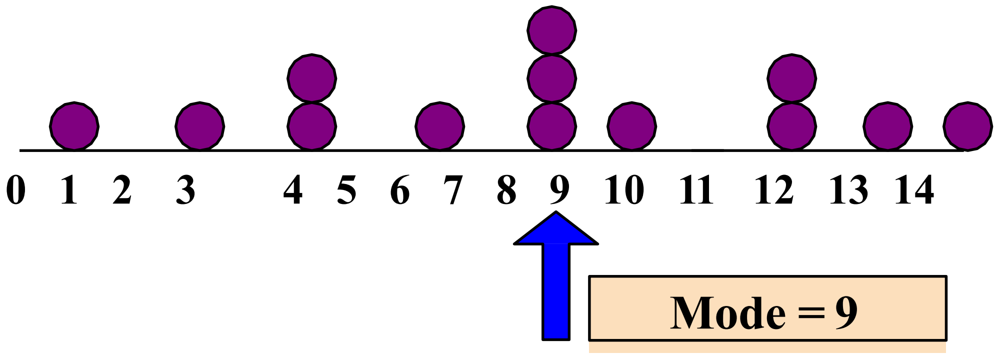
Advantages:

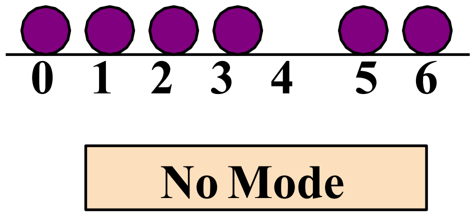
- Not affected extreme values
- Can be computed in case of open class, if median is not in open class
- Can be computed in case categorical variable

- **DisAd:** Arraying of the data is time consuming.
    - To estimate population parameter, mean is easier.

Measures of Central Tendency: The Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either **numerical or categorical** data
- There may be no mode
- There may be several modes





Mode for Grouped Data

$$\text{Mode} = \boxed{L + \frac{d_1}{d_1 + d_2} \times c}$$

Where L =Lower limit of the modal class

$\boxed{d_1 = f_1 - f_0}$ $\boxed{d_2 = f_1 - f_2}$

$\boxed{f_1}$ = Frequency of the **modal class**

$\boxed{f_0}$ = Frequency **preceding** the modal class

$\boxed{f_2}$ = Frequency **succeeding** the modal class. C = **Class Interval** of the modal class

Mode for Grouped Data Example

Example: Find the mode for the following continuous frequency distribution:

Class	0-1	1-2	2-3	3-4	4-5	5-6
Frequency	1	4	8	7	3	2

Solution for the Example

Class	Frequency
0-1	1
1-2	4
2-3	8
3-4	7
4-5	3
5-6	2
Total	**25**

$$\text{Mode} = L + \frac{d_1}{d_1 + d_2} \times c$$

$L = 2$

$d_1 = f_1 - f_0 = 8\text{-}4 = 4$

$d_2 = f_1 - f_2 = 8\text{-}7 = 1$

$C = 1$ Hence $\text{Mode} = 2 + \frac{4}{5} \times 1$

$= 2.8$

Numerical Measures of Central Tendency

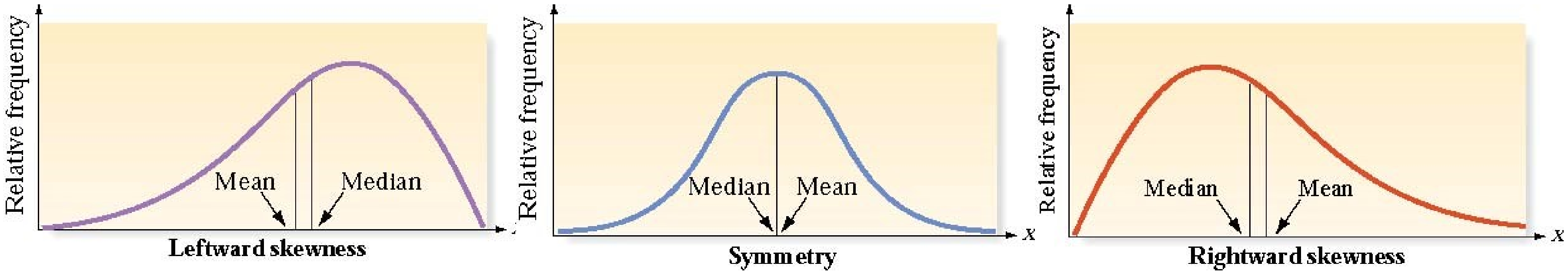
- Perfectly symmetric data set:
  - Mean = Median = Mode
- Extremely high value in the data set:
  - Mean > Median > Mode

    (Rightward skewness)
- Extremely low value in the data set:
  - Mean < Median < Mode

    (Leftward skewness)

Numerical Measures of Central Tendency

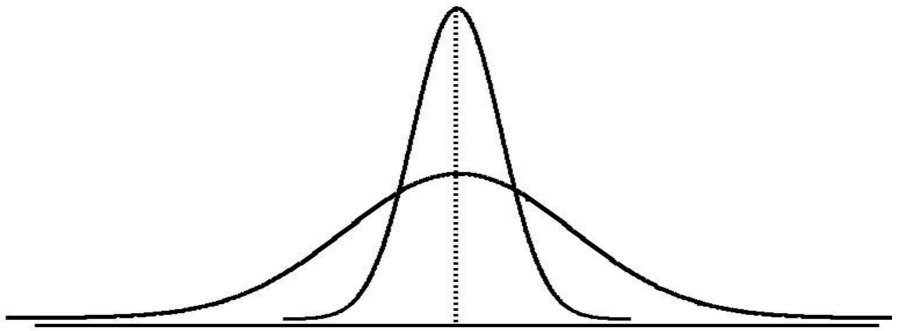
- A data set is **skewed** if one tail of the distribution has more extreme observations than the other tail.



Measures of Variation

- **Variation**
  - **Range**
  - **Variance**
  - **Standard Deviation**
  - **Coefficient of Variation**

- Measures of variation give information on the **spread** or **variability** or **dispersion** of the data values.



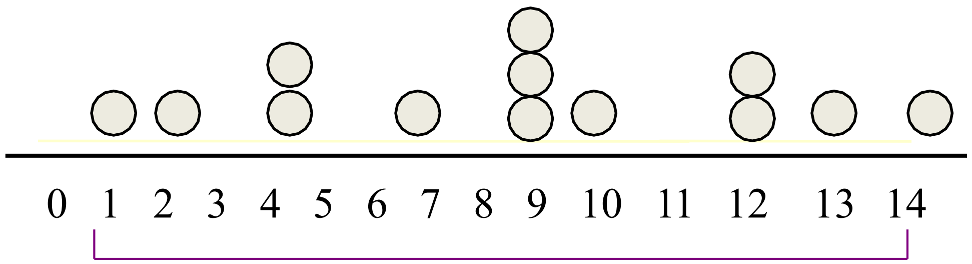
Same center, different variation

Measures of Variation: The Range

- Simplest measure of variation
- Difference between the largest and the smallest values:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

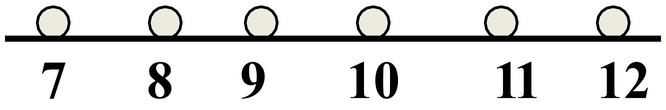
Example:



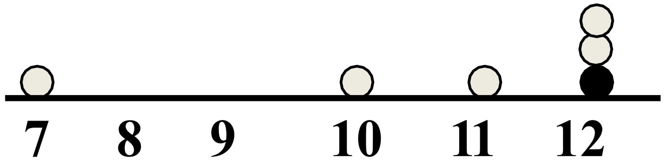
Range = 14 - 1 = 13

Measures of Variation:
Why The Range Can Be Misleading

- **Ignores** the way in which data are **distributed**



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Measures of Variation: The Standard Deviation

- **Most commonly** used measure of variation
- Shows variation about the **mean**
- Is the **square root of the variance**
- Has the same units as the original data

– Sample standard deviation:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Measures of Variation: The Standard Deviation

Steps for Computing Standard Deviation

1. Compute the **difference between each value** and the **mean**.
2. **Square** each difference.
3. **Add the squared** differences.
4. **Divide this total by n-1 to** get the sample variance.
5. Take the **square root of the sample** variance to get the sample standard deviation.

Measures of Variation: Sample Standard Deviation

Activity:

Sample Data (X_i) :	**10 12 14 15 17 18 18 24**

Standard Deviation (Sample) for Grouped Data

Frequency Distribution of Return on Investment of Mutual Funds

Return on Investment	**Number of Mutual Funds**
5-10	10
10-15	12
15-20	16
20-25	14
25-30	8
Total	**60**

Measures of Variation: The Variance

- Average (approximately) of squared deviations of values from the mean

  – Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Where $\bar{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Sample statistics versus population parameters

Measure	Population Parameter	Sample Statistic
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

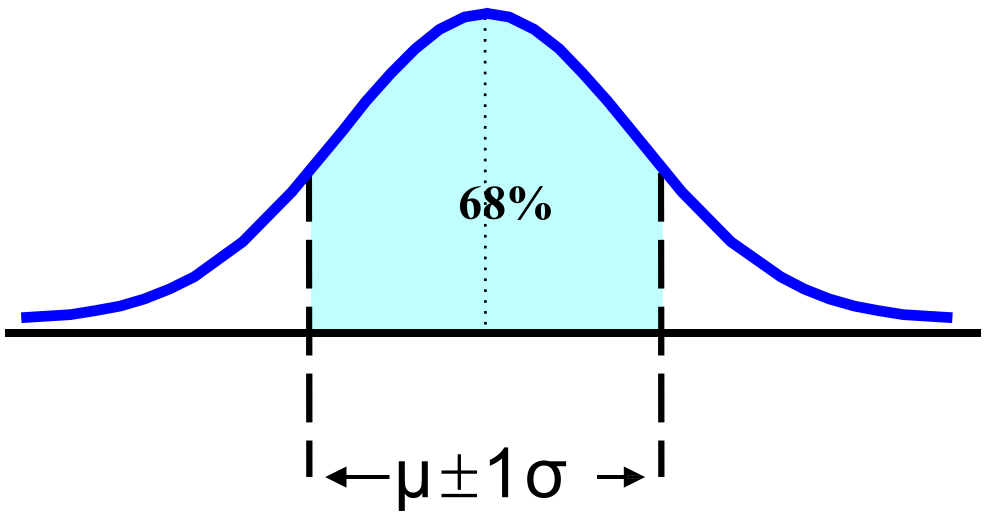
Interpreting the Standard Deviation

- Chebyshev's Rule
- The Empirical Rule

Both tell us something about where the data will be relative to the mean.

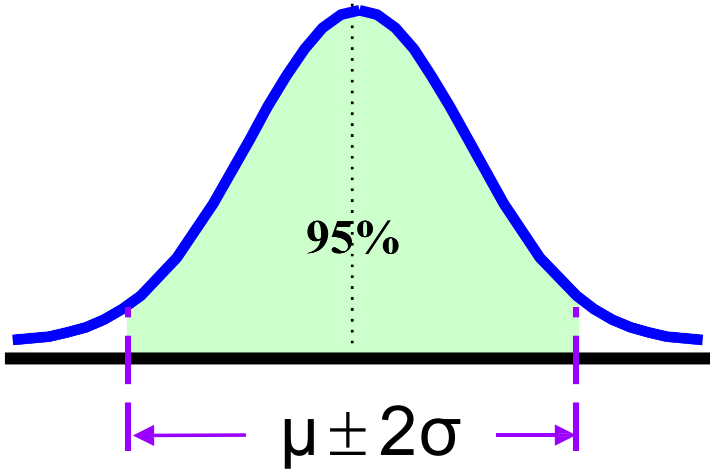
Numerical Descriptive Measures: The Empirical Rule for distribution of data

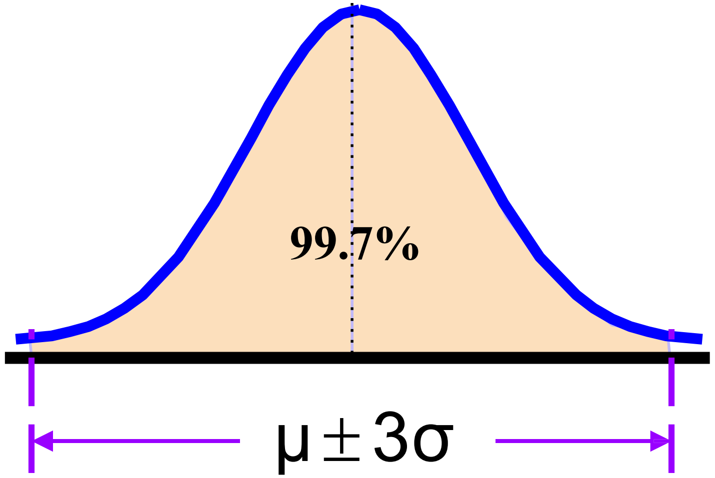
- The empirical rule approximates the variation of data in **a bell-shaped** distribution
- Approximately 68. 27% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\mu \pm 1\sigma$



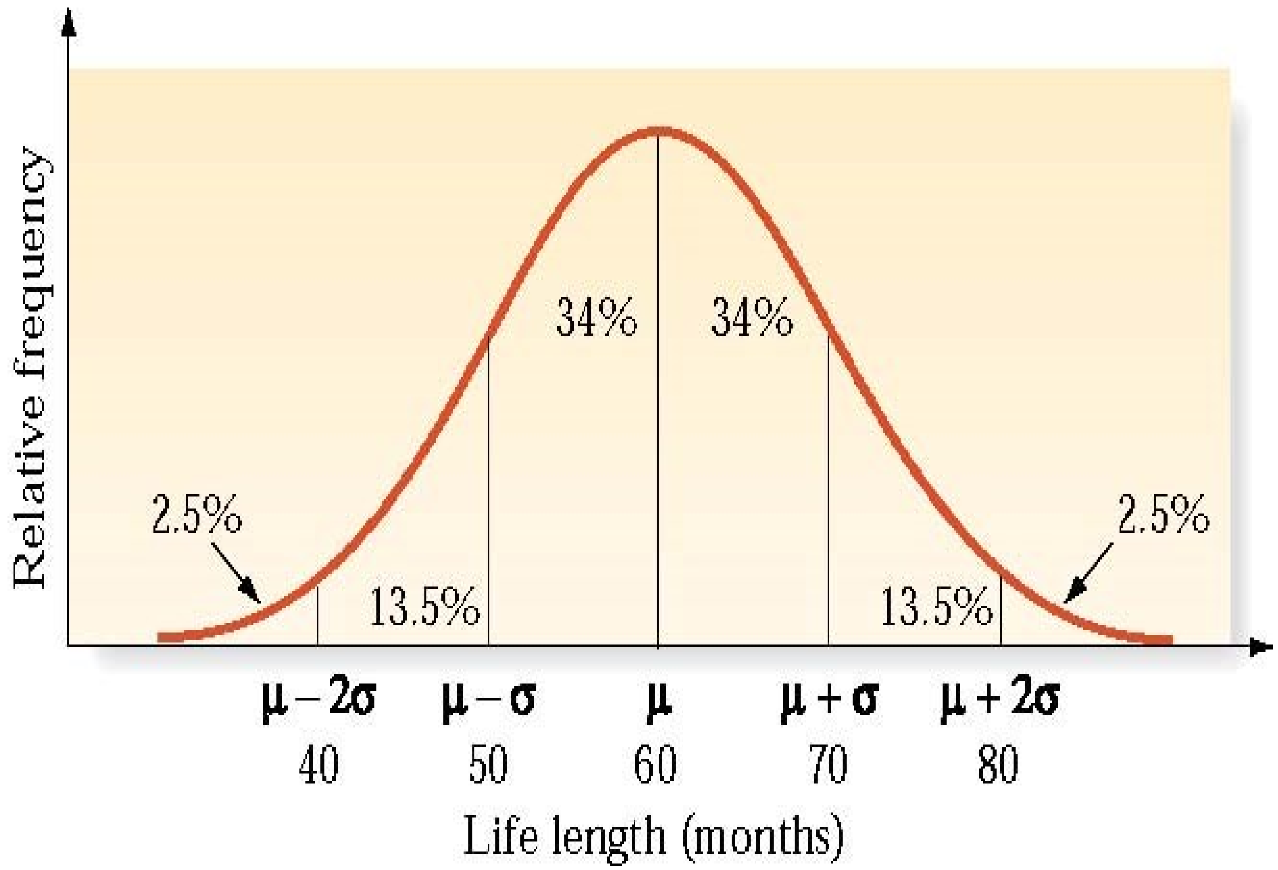
The Empirical Rule

- Approximately 95.45% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$
- Approximately 99.73% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$





Interpreting the Standard Deviation



- The Empirical Rule
  - Useful for mound-shaped, symmetrical distributions
  - If not perfectly mounded and symmetrical, the values are approximations
- For a perfectly symmetrical and mound-shaped distribution,
  - ~68% will be within the range $(\bar{x}-s, \bar{x}+s)$
  - ~95% will be within the range $(\bar{x}-2s, \bar{x}+2s)$
  - ~99.7% will be within the range $(\bar{x}-3s, \bar{x}+3s)$

Interpreting the Standard Deviation

- Chebyshev's Rule
  - Valid for *any* data set
  - For any number $k > 1$, at least $(1-1/k^2)\%$ of the observations will lie within k standard deviations of the mean

k	k^2	$1/k^2$	$(1- 1/k^2)\%$
2	4	.25	75%
3	9	.11	89%
4	16	.0625	93.75%

Interpreting the Standard Deviation

- How many observations fit within $\pm$ *n* s of the mean?

	Chebyshev's Rule	Empirical Rule
$\pm 1s$ or $\pm 1\sigma$	No useful info	Approximately 68%
$\pm 2s$ or $\pm 2\sigma$	At least 75%	Approximately 95%
$\pm 3s$ or $\pm 3\sigma$	At least 8/9	Approximately 99.7%

Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

Since 45 and 65 are exactly one standard deviation below and above the mean, the empirical rule says that about 68% of the hummingbirds will be in this range.

Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
    - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
    - Between 55 and 65?
    - Less than 45?

This range of numbers is from the mean to one standard deviation above it, or one-half of the range in the previous question. So, about one-half of 68%, or 34%, of the hummingbirds will be in this range.

Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.
  - Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
  - Between 55 and 65?
  - Less than 45?

Half of the entire data set lies above the mean, and ~34% lie between 45 and 55 (between one standard deviation below the mean and the mean), so ~84% (~34% + 50%) are above 45, which means ~16% are below 45.

Interpreting the Standard Deviation

- You have purchased compact fluorescent light bulbs for your home. Average life length is 500 hours, standard deviation is 24, and frequency distribution for the life length is mound shaped. One of your bulbs burns out at 450 hours. Would you send the bulb back for a refund?

Interval	Range	% of observations included	% of observations excluded
$\pm 1s$	476 - 524	Approximately 68%	Approximately 32%
$\pm 2s$	452 - 548	Approximately 95%	Approximately 5%
$\pm 3s$	428 - 572	Approximately 99.7%	Approximately 0.3%

Measures of Variation: Comparing Standard Deviations

The **coefficient of variation** (CV) is a measure of relative **variability**.

It is the ratio of the **standard deviation to the mean** (average).

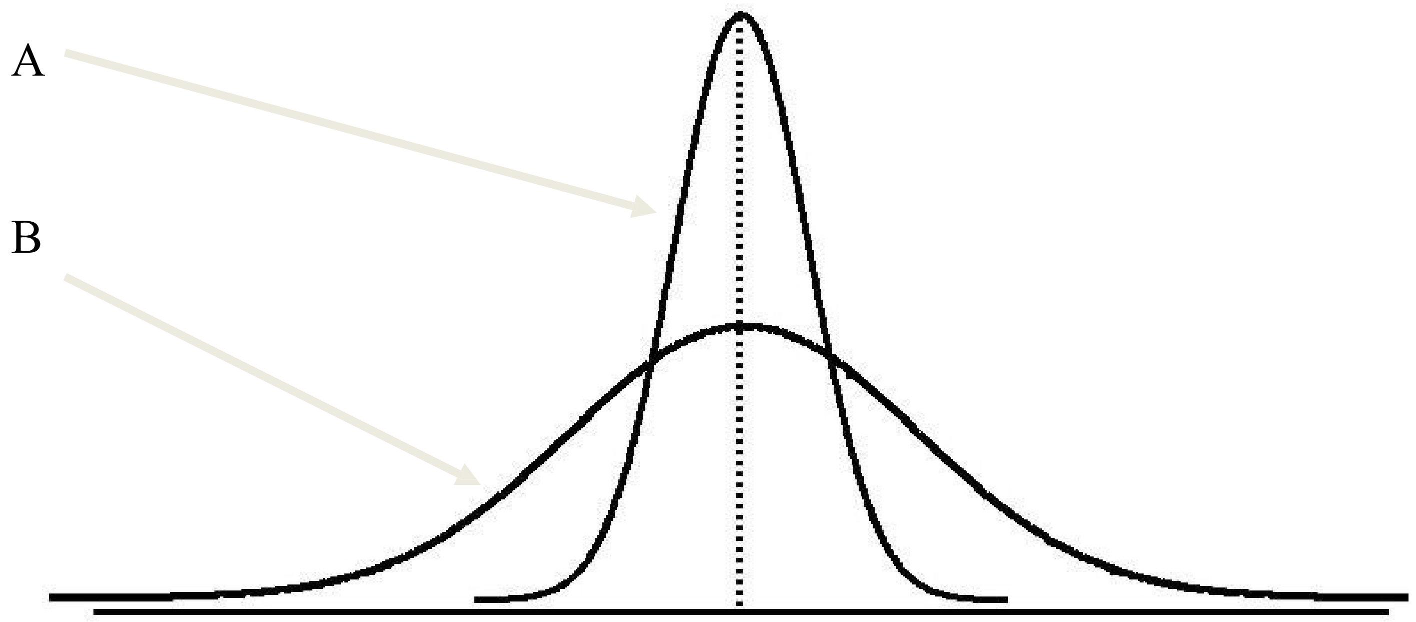
Always in percentage (%)

Shows **variation relative to mean**

Can be used to compare the variability of two or more sets of data measured in **different units**

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

Measures of Variation:
Comparing Standard Deviations



Which curve
has higher SD?

Measures of Variation: Comparing Coefficients of Variation

- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = \$100
  - Standard deviation = \$5

$$CV_B = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Numerical Measures of Relative Standing

- Descriptive measures of relationship of a measurement to the rest of the data
- Common measures:
    - percentile ranking or percentile score
    - z-score

Numerical Measures of Relative Standing

- The *z-score* tells us how many standard deviations above or below the mean a particular measurement is.

- Sample z-score

$$z = \frac{x - \overline{x}}{s}$$

- Population z-score

$$z = \frac{x - \mu}{\sigma}$$

Z-Score

- To compute the Z**-score (Standard score)** of a data value, **subtract the mean** and **divide by the standard deviation.**

- **The Z-score is the number of standard deviations a data value is from the mean.**

- A data value is considered an extreme outlier if its Z-score is **less than -3.0 or greater than +3.0.**

- The **larger** the absolute value of the Z-score, the **farther** the data value is from the mean.

Locating Extreme Outliers: Z-Score

- Suppose the **mean** math SAT score is 490, with a standard deviation of 100.
- Compute the Z-score for a test score of 620.

$$Z = \frac{X - \bar{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

A score of 620 is 1.3 standard deviations above the mean and would **not be considered an outlier**.

Interpreting the Standard Deviation

- Hummingbirds beat their wings in flight an average of 55 times per second.
- Assume the standard deviation is 10, and that the distribution is symmetrical and mounded.

$$z = \frac{x - \overline{x}}{s}$$

An individual hummingbird is measured with 75 beats per second. What is this bird's z-score?

$$z = \frac{75 - 55}{10} = 2.0$$

Numerical Measures of Relative Standing

- Z scores are related to the empirical rule:

  For a perfectly symmetrical and mound-shaped distribution,

  - ~68 % will have z-scores between -1 and 1
  - ~95 % will have z-scores between -2 and 2
  - ~99.7% will have z-scores between -3 and 3

Methods for Determining Outliers

- Outliers and z-scores
  - The chance that a z-score is between -3 and +3 is over 99%.

  - Any measurement with $|z| > 3$ is considered an outlier.

xi	xi-x	(xi-x)/SD
240	-140	-1.237437797
260	-120	-1.060660969
350	-30	-0.265165242
350	-30	-0.265165242
420	40	0.353553656
510	130	1.149049383
530	150	1.325826211
Mean 380		

SD= 113

Is there any outlier???

Numerical Measures of Relative Standing

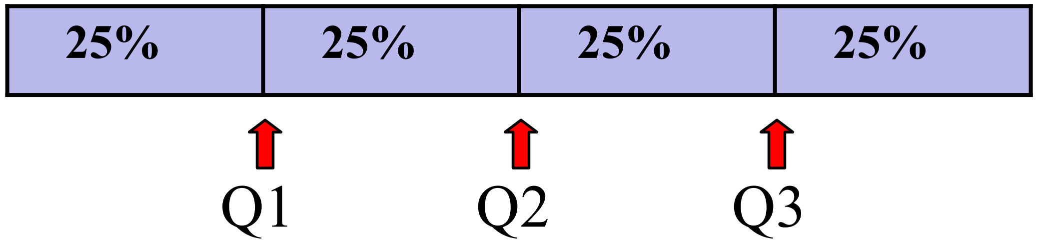
- **Percentiles**: for any (large) set of *n* measurements (arranged in ascending or descending order), the p^{th} *percentile* is a number such that p% of the measurements fall below that number and $(100 – p)$% fall above it.

Numerical Measures of Relative Standing

- Finding percentiles is similar to finding the median – the median is the 50^{th} percentile.
  - If you are in the 50^{th} percentile for the GRE, half of the test-takers scored better and half scored worse than you.
  - If you are in the 75^{th} percentile, you scored better than three-quarters of the test-takers.
  - If you are in the 90^{th} percentile, only 10% of all the test-takers scored better than you.

Quartiles

Quartiles split the ranked **data into 4 segments with** an equal number of values per segment



- The first quartile, Q_1, is the value for which **25% of the observations are smaller** and 75% are larger

- Q_2 is the same as the median (50% of the observations are smaller and 50% are larger)

- Only 25% of the observations are **greater than the third quartile**

Locating Quartiles

Find a quartile by determining the value in the appropriate **position** in the ranked data, where

First quartile position: $Q1 = (n+1)/4$ ranked value

Second quartile position: $Q2$ = Same procedure as median

Third quartile position: $Q3 = 3(n+1)/4$ ranked value

- where n is the number of observed values

Locating Quartiles

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data

so use the value half way between the 2nd and 3rd values,

so $\boxed{Q_1 = 12.5}$

Q_1 and Q_3 are measures of **non-central location**
Q_2 = median, is a measure of **central tendency**

Quartile Example

Sample Data in Ordered Array: 11 12 13 16 16 17 18 21 22

$(n = 9)$

Q_1 is in the $(9+1)/4 = 2.5$ position of the ranked data,

so $\mathbf{Q_1 = (12+13)/2 = 12.5}$

Q_2 is in the $(9+1)/2 = 5^{th}$ position of the ranked data,

so $\mathbf{Q_2 = \text{median} = 16}$

Q_3 is in the $3(9+1)/4 = 7.5$ position of the ranked data,

so $\mathbf{Q_3 = (18+21)/2 = 19.5}$

Quartile Measures: The Interquartile Range (IQR)

- The IQR is $Q_3 - Q_1$ and measures the spread in the **middle 50% of** the data

- The IQR is also called the **midspread** because it covers the middle 50% of the data

- The IQR is ***a measure of variability that is not influenced by*** **<u>outliers</u>** or extreme values

- Measures like Q_1, Q_3, and IQR that are ***not influenced by outliers are called <u>resistant measures</u>***

The Five Number Summary

The five numbers that help describe the **center, spread and shape of data are:**

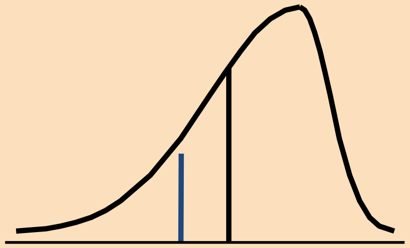
- $X_{smallest}$
- First Quartile (Q_1)
- Median (Q_2)
- Third Quartile (Q_3)
- $X_{largest}$

Relationships among the five-number summary and distribution shape

Left-Skewed	Symmetric	Right-Skewed
$\text{Median} - X_{\text{smallest}}$ $>$ $X_{\text{largest}} - \text{Median}$	$\text{Median} - X_{\text{smallest}}$ $\approx$ $X_{\text{largest}} - \text{Median}$	$\text{Median} - X_{\text{smallest}}$ $<$ $X_{\text{largest}} - \text{Median}$
$Q_1 - X_{\text{smallest}}$ $>$ $X_{\text{largest}} - Q_3$	$Q_1 - X_{\text{smallest}}$ $\approx$ $X_{\text{largest}} - Q_3$	$Q_1 - X_{\text{smallest}}$ $<$ $X_{\text{largest}} - Q_3$
$\text{Median} - Q_1$ $>$ $Q_3 - \text{Median}$	$\text{Median} - Q_1$ $\approx$ $Q_3 - \text{Median}$	$\text{Median} - Q_1$ $<$ $Q_3 - \text{Median}$

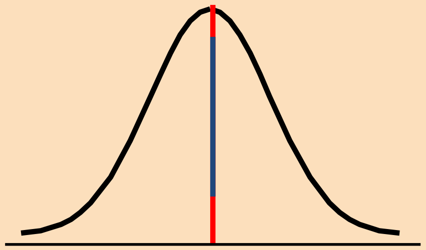
Left-Skewed(Negative))

Mean < Median



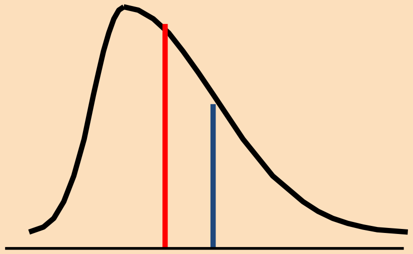
Symmetric

Mean = Median



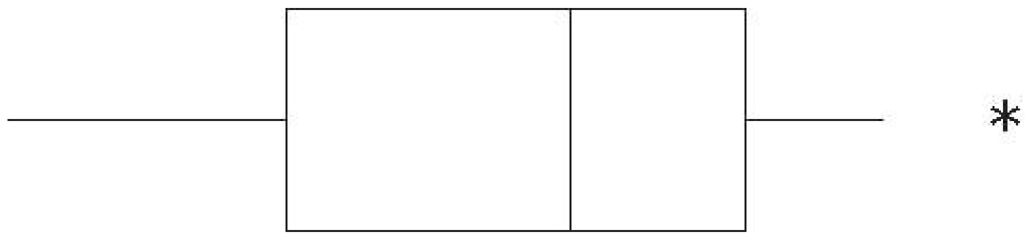
Right-Skewed (Positive)

Median < Mean



The Box Plot

- The **box plot** is a graph representing information about certain percentiles for a data set and can be used to identify outliers

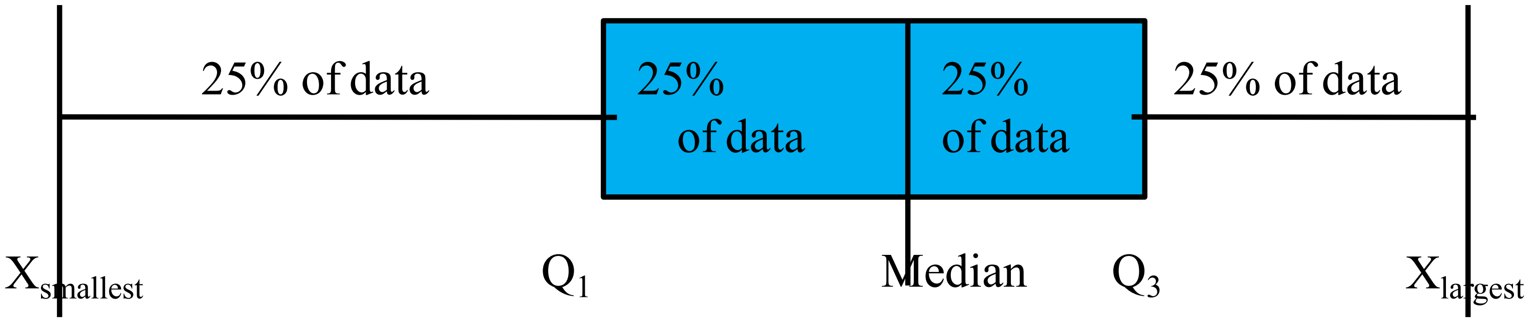


Five Number Summary and The Boxplot

- The Boxplot: A Graphical display of the data based on the five-number summary:

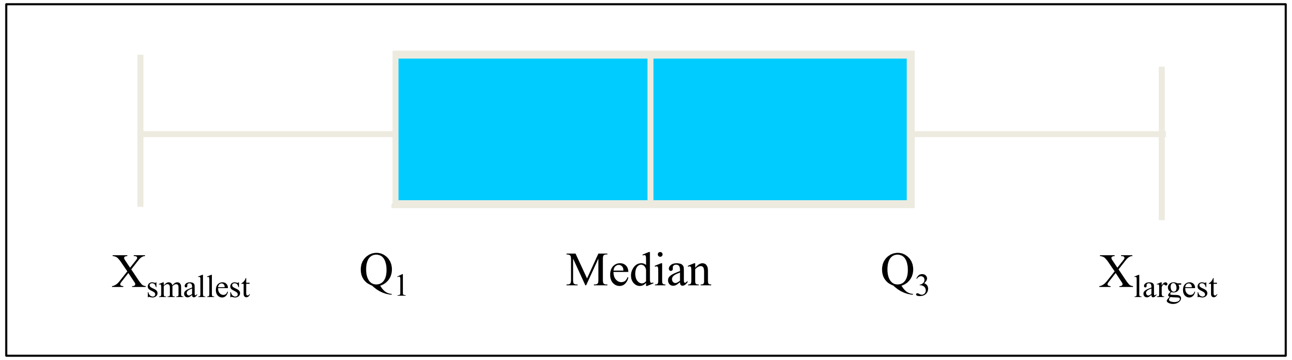
$X_{smallest}$ -- Q_1 -- Median -- Q_3 -- $X_{largest}$

Example:



Five Number Summary: Shape of Boxplots

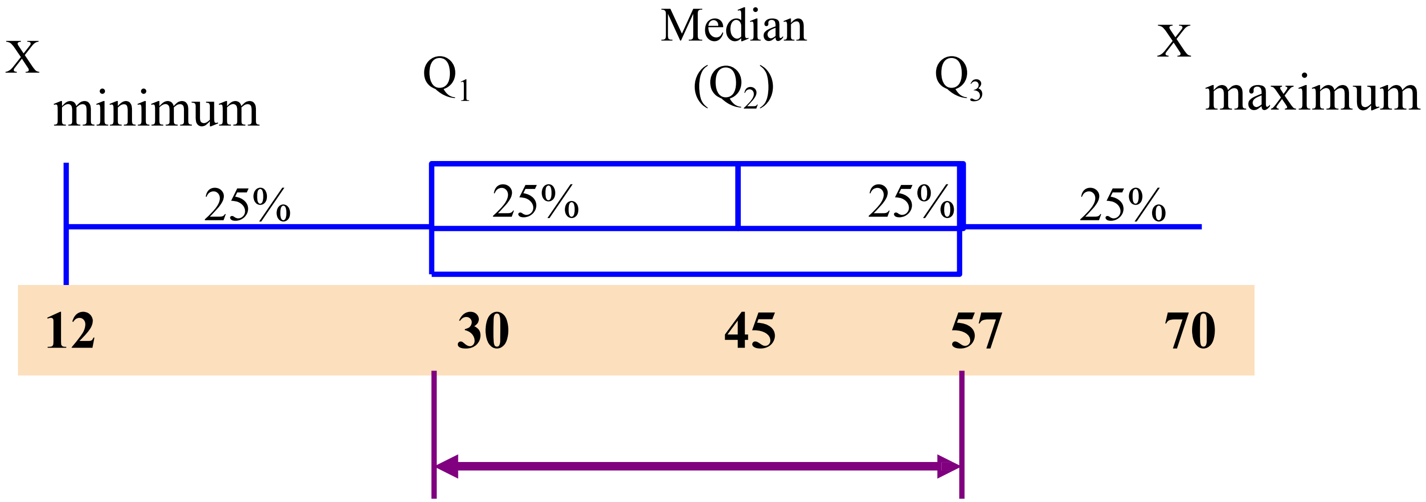
- If data **are symmetric around** the median then the box and central line are **centered between the endpoints**



- A Boxplot can be shown in either a **vertical or horizontal** orientation

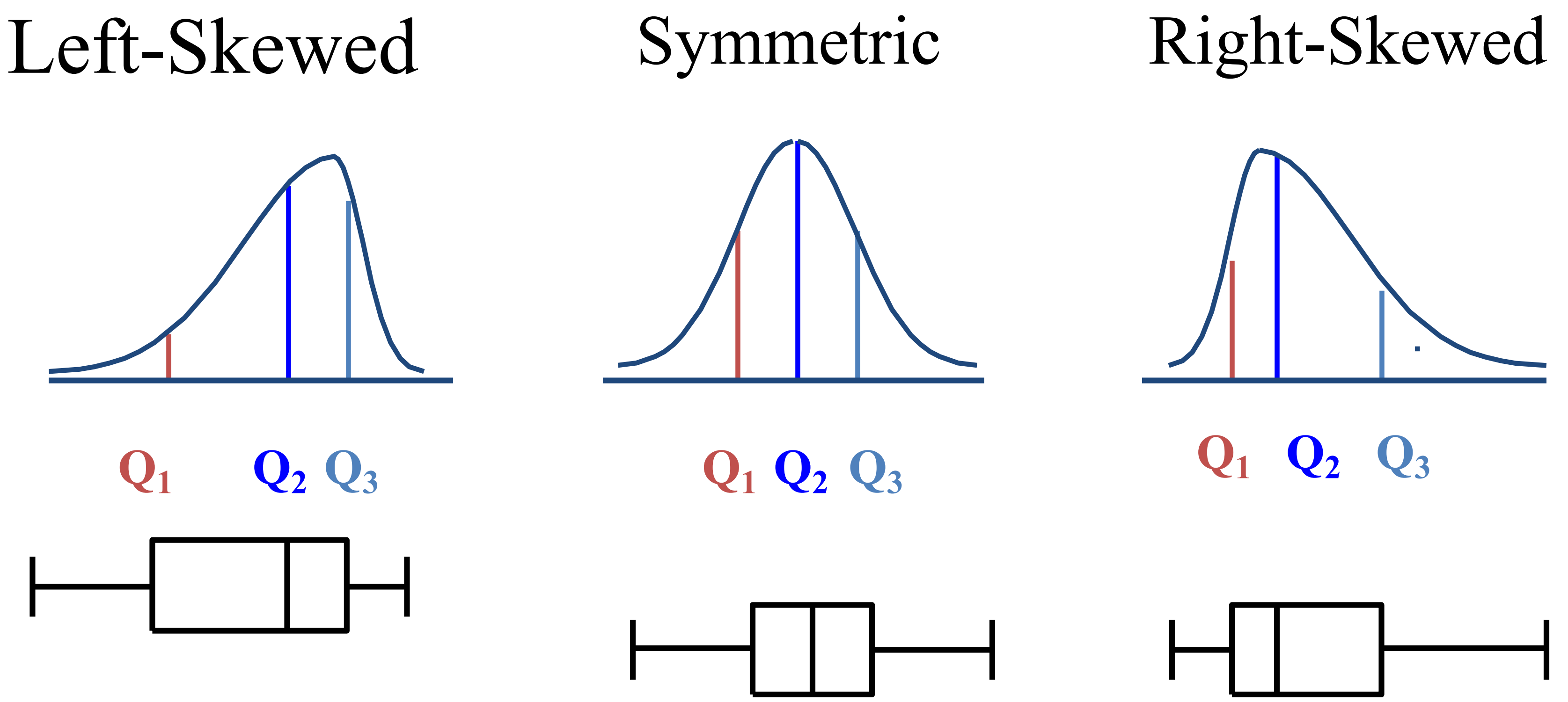
Calculating The Interquartile Range

Example:



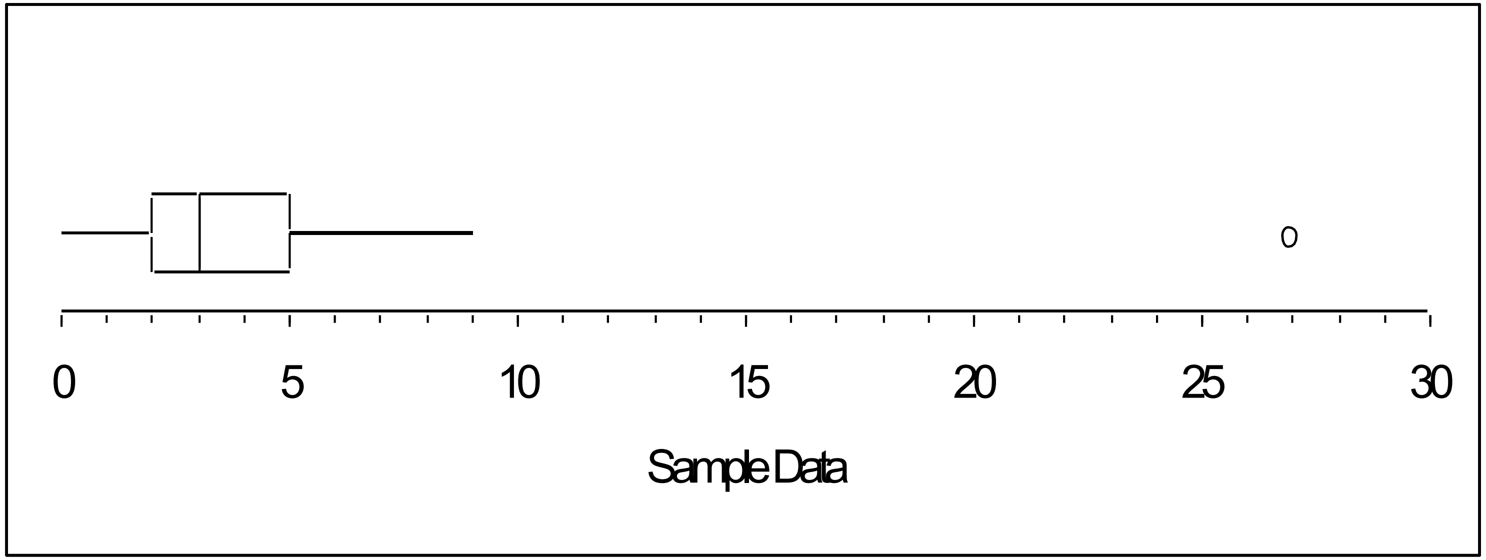
Interquartile range
$= 57 - 30 = 27$

Distribution Shape and The Boxplot

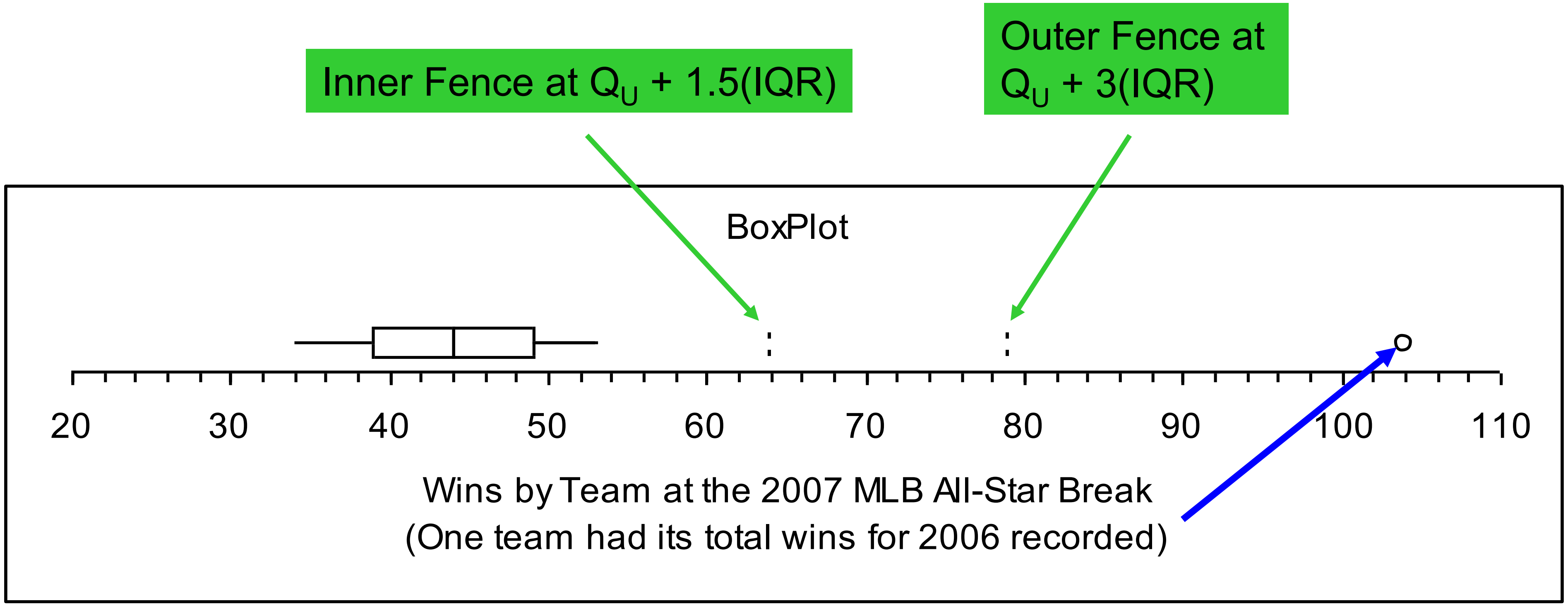


Boxplot example showing an outlier

- The boxplot below of the same data shows the outlier **value of 27 plotted separately**
- A value is considered an outlier if it is **more than 1.5 times** the interquartile range **below Q_1** or **above Q_3**



Methods for Determining Outliers



Methods for Detecting Outliers

- Rules of thumb
- Box Plots
  - measurements between inner and outer fences are suspect
  - measurements beyond outer fences are highly suspect
- Z-scores
  - Scores of ± 3 in mounded distributions (± 2 in highly skewed distributions) are considered outliers

Find outliers?

850	875	4700	4900	5300	5700	6700	7300	7700	8100
8300	8400	8700	8700	8900	9300	9500	9500	9700	10000
10300	10500	10700	10800	11000	11300	11300	11800	12700	12900
13100	13500	13800	14900	16300	17200	18500	20300	21310	21315

Relationship between <u>two numerical</u> variable

1 Covariance

2 Coefficient of correlation

- The **covariance** measures the **strength of the linear** relationship between **two numerical variables** (X & Y)

- The sample covariance:

$$\boxed{\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}}$$

- Only concerned with the ***<u>strength of the relationship</u>***
- **No causal effect** is implied

Interpreting Covariance

- **Covariance** between two variables:

$\text{cov}(X,Y) > 0 \longrightarrow$ X and Y tend to move in the same direction

$\text{cov}(X,Y) < 0 \longrightarrow$ X and Y tend to move in opposite directions

$\text{cov}(X,Y) = 0 \longrightarrow$ X and Y are independent

- The covariance has a major flaw:
  - It is not possible to determine the ***relative strength of the relationship*** from the **size** of the covariance

Find covariance??

Sr no	City	Hamburger (x)	Movie Tickets (y)
1	Tokyo	5.99	32.66
2	London	7.62	28.41
3	New York	5.75	20.00
4	Sydney	4.45	20.71
5	Chicago	4.99	18.00
6	San Francisco	5.29	19.50
7	Boston	4.39	18.00
8	Atlanta	3.7	16.00
9	Toronto	4.62	18.05
10	Rio de Janeiro	2.99	9.90
Avg		4.98	20.12

Sr no	City	Hamburger (x)	Movie Tickets (y)	(x-x bar)*(y-ybar)
1	Tokyo	5.99	32.66	12.6654
2	London	7.62	28.41	21.8856
3	New York	5.75	20.00	-0.0924
4	Sydney	4.45	20.71	-0.3127
5	Chicago	4.99	18.00	-0.0212
6	San Francisco	5.29	19.50	-0.1922
7	Boston	4.39	18.00	1.2508
8	Atlanta	3.7	16.00	5.2736
9	Toronto	4.62	18.05	0.7452
10	Rio de Janeiro	2.99	9.90	20.3378
Avg		4.98	20.12	Sum= 61.53

Covariance = 61.53/9=6.83, we can’t tell whether this value is an indictor of strong or weak relationship.

Coefficient of Correlation

- Measures the relative strength of the linear relationship between two numerical variables
- Sample coefficient of correlation:

$$\boxed{r = \frac{\text{cov}(X, Y)}{S_X S_Y}}$$

where

$$\boxed{\text{cov}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}$$

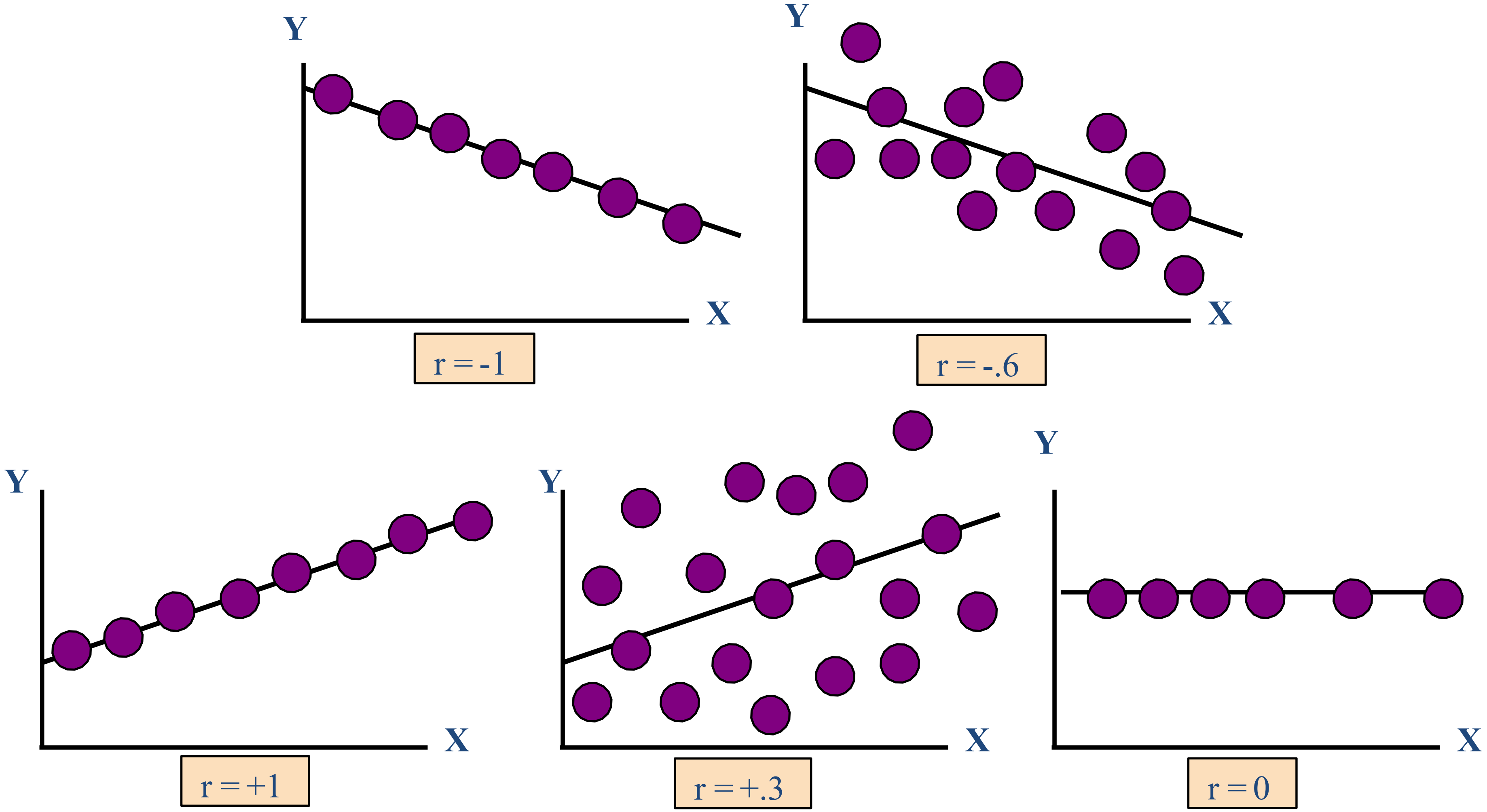
$$\boxed{S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1}}}$$

$$\boxed{S_Y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}}$$

Features of the Coefficient of Correlation

- The **population** coefficient of correlation is referred as **ρ**.
- The **sample** coefficient of correlation is referred to as **r**.
- Either ρ or r have the following **features**:
  - Unit free
  - Ranges between **–1 and 1**
  - The closer to –1, the stronger the **negative linear** relationship
  - The closer to 1, the stronger the **positive linear** relationship
  - The closer to 0, the **weaker** the **linear** relationship

Scatter Plots of Sample Data with Various Coefficients of Correlation



Product	Calories	Fat
Dunkin' Donuts Iced Mocha Swirl latte (whole milk)	240	8
Starbucks Coffee Frappuccino blended coffee	260	3.5
Dunkin' Donuts Coffee Coolatta (cream)	350	22
Starbucks Iced Coffee Mocha Expresso (whole milk and whipped cream	350	20
Starbucks Mocha Frappuccino blended coffee (whipped cream)	420	16
Starbucks Chocolate Brownie Frappuccino blended coffee (whipped cream)	510	22
Starbucks Chocolate Frappuccino Blended Crème (whipped cream)	530	19

a) Compute covariance
b) Compute coefficient of correlation
c) Which is valuable in expressing relationship
d) What conclusion can you reach about relationship

Product	Calories	Fat
Dunkin' Donuts Iced Mocha Swirl latte (whole milk)	240	8
Starbucks Coffee Frappuccino blended coffee	260	3.5
Dunkin' Donuts Coffee Coolatta (cream)	350	22
Starbucks Iced Coffee Mocha Expresso (whole milk and whipped cream	350	20
Starbucks Mocha Frappuccino blended coffee (whipped cream)	420	16
Starbucks Chocolate Brownie Frappuccino blended coffee (whipped cream)	510	22
Starbucks Chocolate Frappuccino Blended Crème (whipped cream)	530	19

a) Compute covariance : 591.66
b) Compute coefficient of correlation: r = 0.71
c) Which is valuable in expressing relationship: **correlation**
d) What conclusion can you reach about relationship: strong positive relationship

Pitfalls in Numerical Descriptive Measures

- Data **analysis** is **objective**
  - Should report the summary measures that best describe and communicate the **important aspects of** the data set

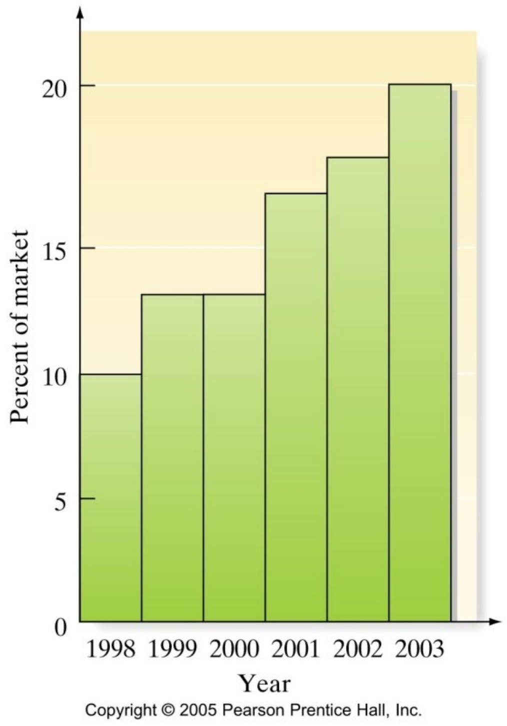
- Data **interpretation is subjective**
  - Should be done **in fair, neutral and clear manner**

Distorting the Truth with Deceptive Statistics

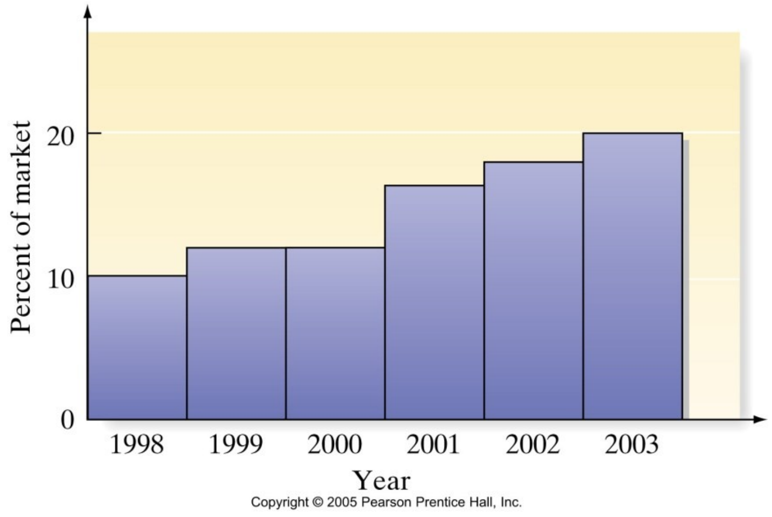
- Distortions
  - Stretching the axis (and the truth)
  - Is average average?
    - Mean, median or mode?
  - Is average relevant?
    - What about the spread?

Distorting the Truth with Descriptive Techniques

- Graphical techniques
  - Scale manipulation

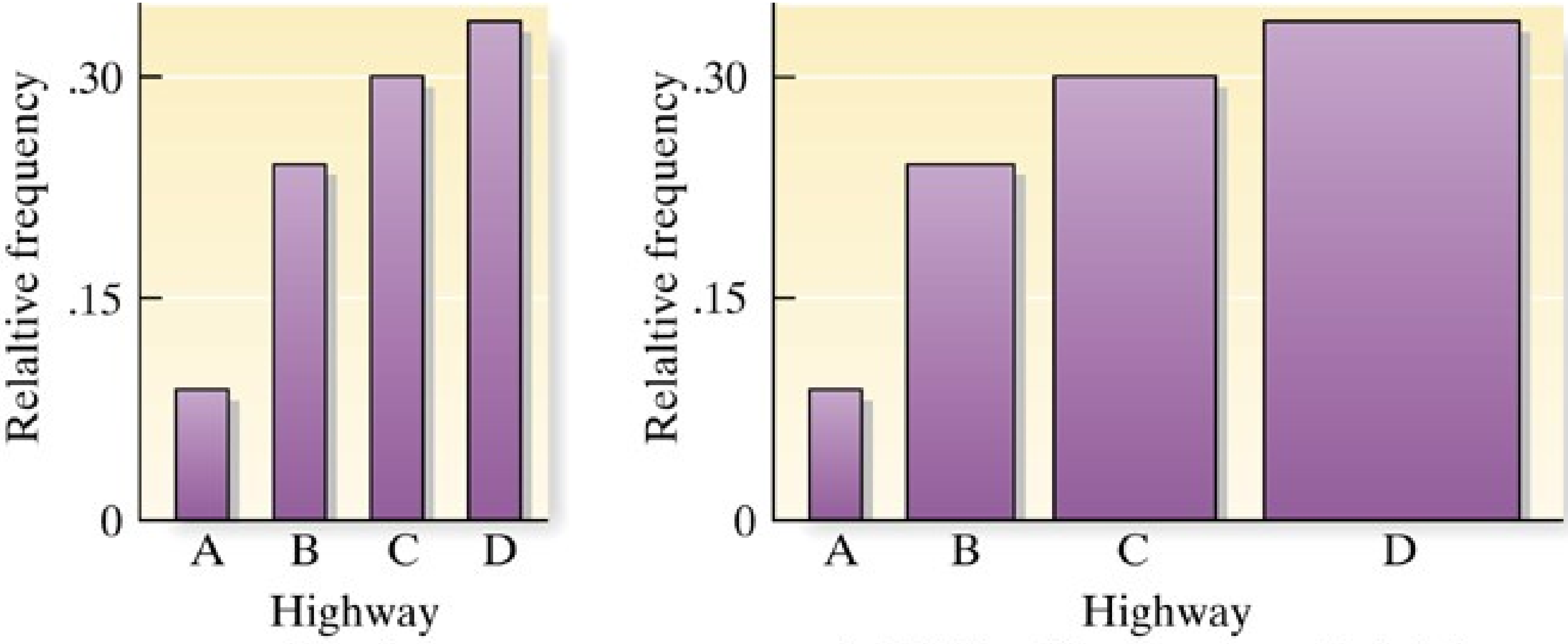


Same data, different scales



Distorting the Truth with Descriptive Techniques

- Graphical techniques
  - More Scale manipulation

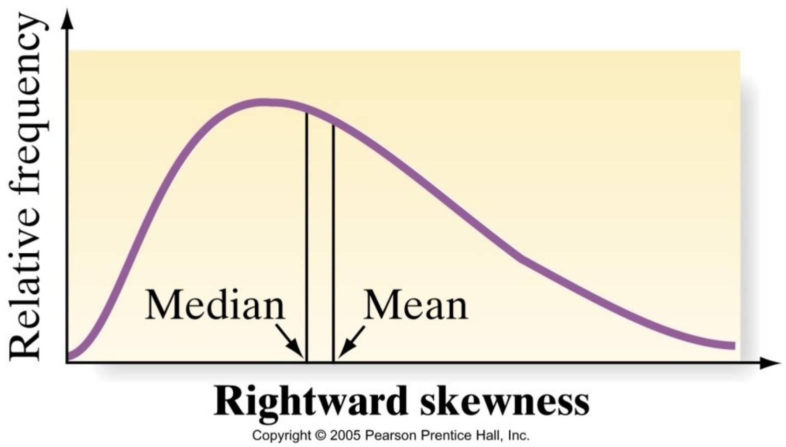


a. Bar chart

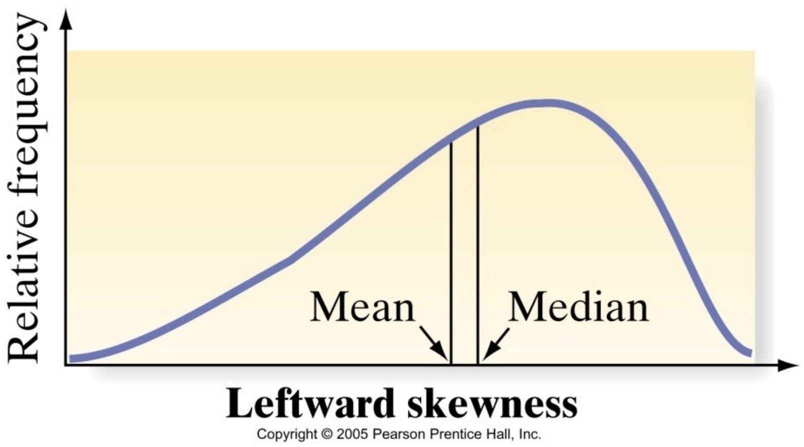
b. Width of bars grows with height

Distorting the Truth with Descriptive Techniques

- Numerical techniques
  - Mismatch of measure of central tendency and distribution shape



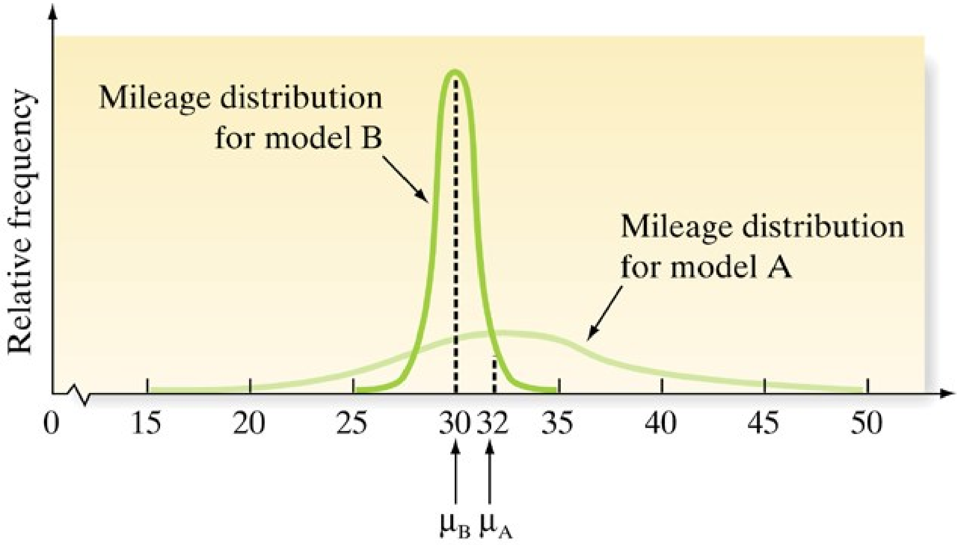
Use of mean overstates average



Use of mean understates average

Distorting the Truth with Descriptive Techniques

- Numerical techniques
  - Discussion of central tendency with no information on variability



Which model would you purchase if you knew only the average MPG?

Would knowing the standard deviation affect your choice?

Why?

Distorting the Truth with Descriptive Techniques

- Graphical techniques
  - Look past the pictures to the data they represent
- Numerical techniques
  - Is measure being used most appropriate for underlying distribution?
  - Are you provided with information on central tendency and variability?

Activity:

X Distribution Company, a subsidiary of a major appliance manufacturer, is forecasting regional sales for next year. The Atlantic branch, with current yearly sales of $193.8 million, is expected to achieve a sales growth of 7.25%; the Midwest branch, with current sales of $79.3 million is expected to grow by 8.20%; and the Pacific branch, with sales of $57.5 million, is expected to increase sales by 7.15%. What is the average rate of sales growth forecasted for next year?

Activity: Talent, Ltd. a Hollywood casting company, is selecting a group of extras for a movie. The ages of the first 20 men to be interviewed are

50	56	55	49	52	57	56	57	56	59
54	55	61	60	51	59	62	52	54	49

The director of the movie wants men whose ages are fairly tightly grouped around **55 years**. The director suggests that a standard deviation of 3 years would be acceptable. Does this group of extras qualify?