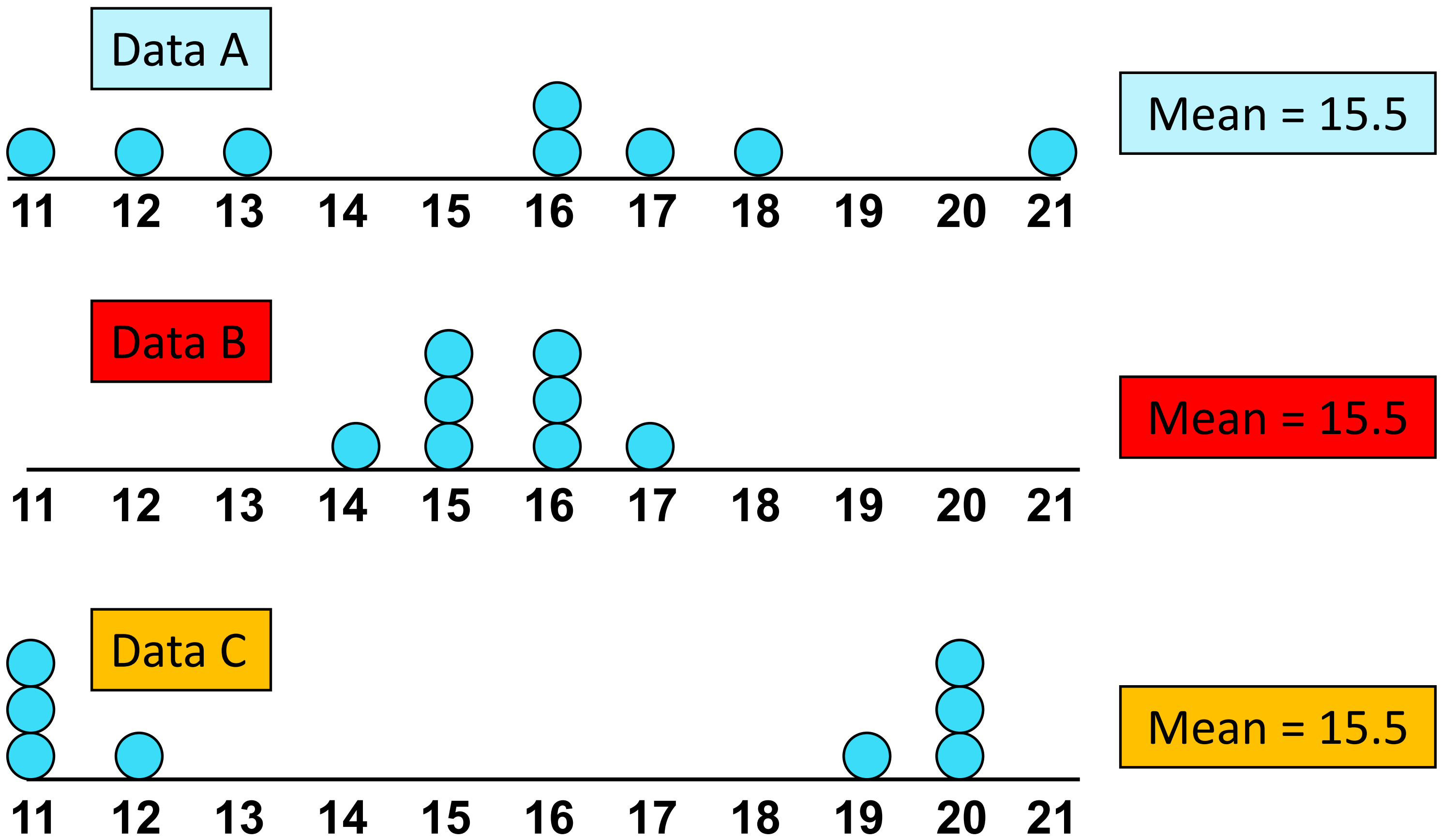
MODULE ONE: PRESENTING AND DESCRIBING INFORMATION

TOPIC 3: NUMERICAL DESCRIPTIVE MEASURES

Numerical DESCRIPTIVE Summary Measures

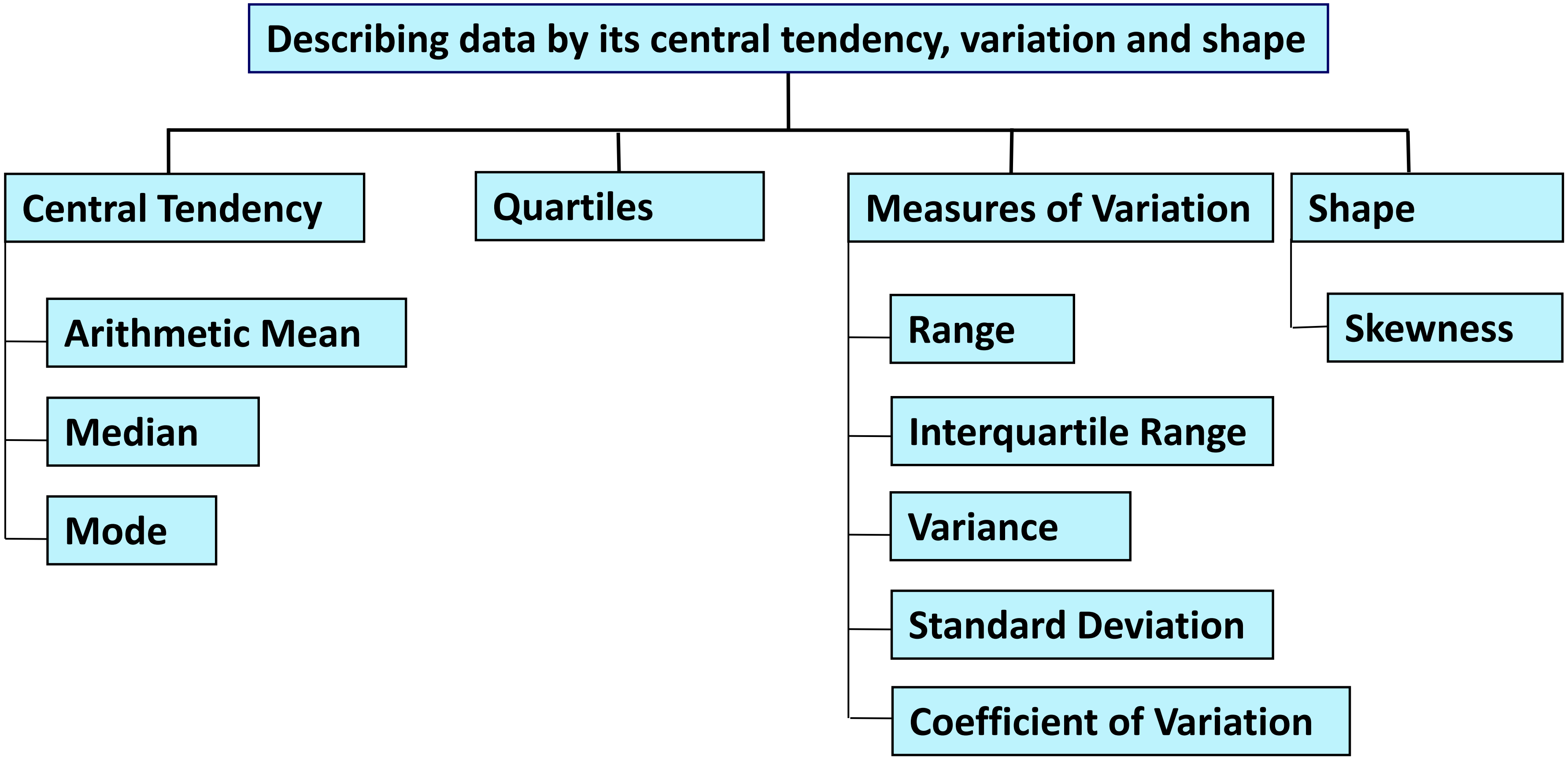


Learning Objectives

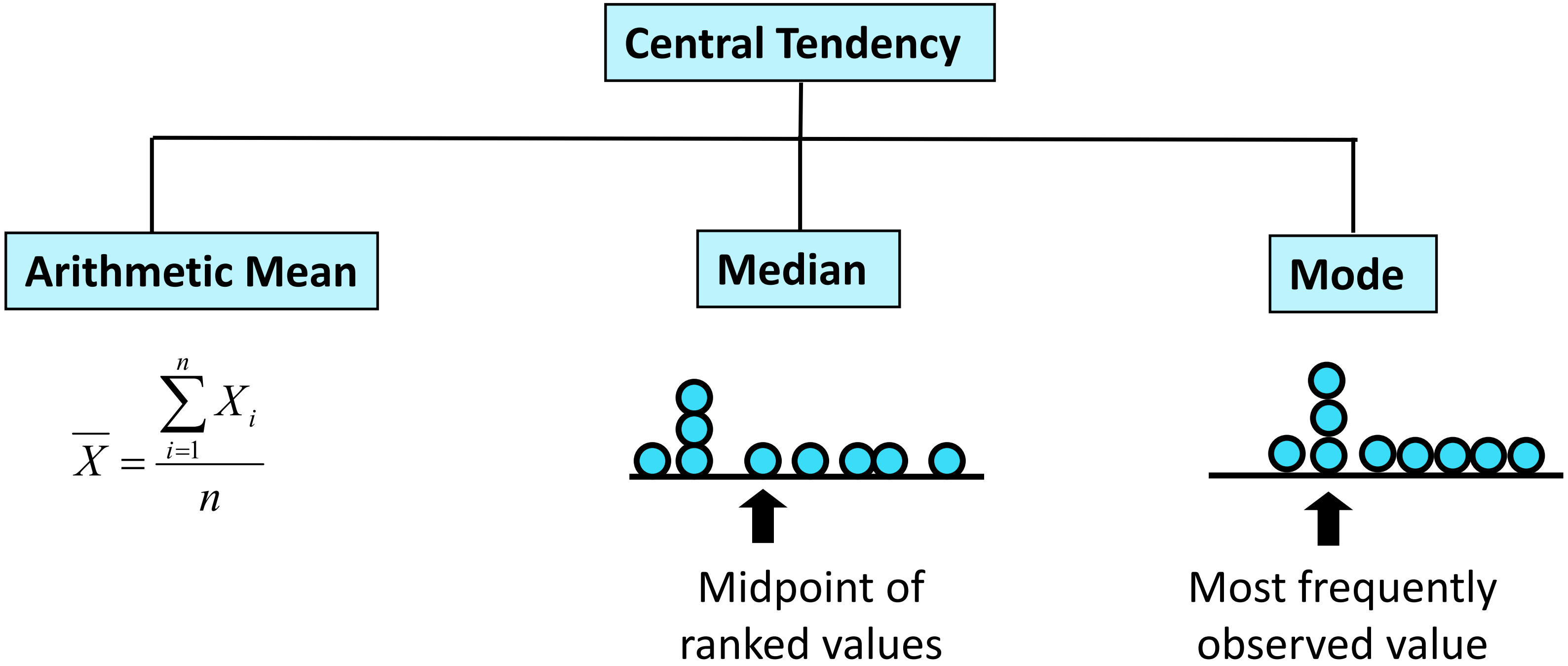
At the completion of this topic, you should be able to:

- calculate and interpret numerical descriptive measures of central tendency, variation and shape for numerical data
- calculate and interpret descriptive summary measures for a population
- describe the relationship between two categorical variables using contingency tables
- describe the relationship between two numerical variables using scatter diagrams and time-series plots
- construct and interpret a box-and-whisker plot
- calculate and interpret the covariance and the coefficient of correlation for bivariate data

+3.1 Measures of Central Tendency, Variation and Shape



+Measures of Central Tendency



$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

+Arithmetic Mean

For a sample of size *n*, the sample mean, denoted $\overline{X}$, is calculated:

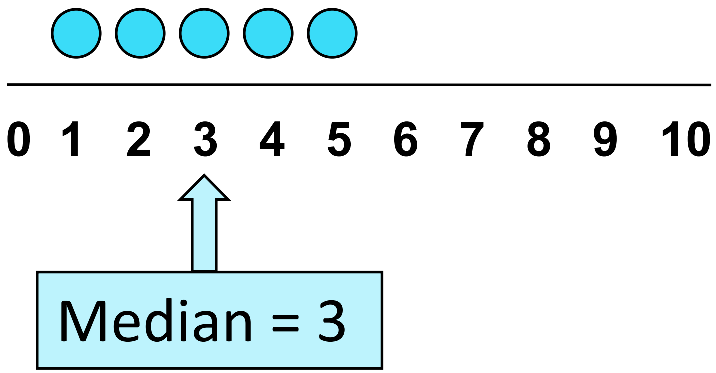
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

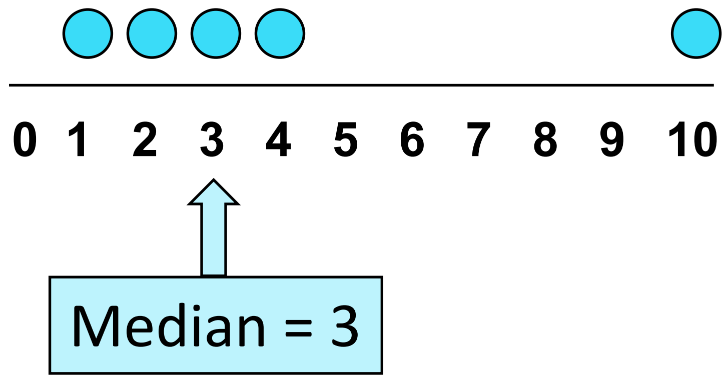
X_i's are observed values

Where Σ means to sum or add up

+Median

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





Its main advantage over the arithmetic mean is that it is not affected by extreme values

+Median

The location of the median:

Median = $\frac{n+1}{2}$ ranked value

- Note that $\frac{n+1}{2}$ is not the **value** of the median, only the **position** of the median in the ranked data

Rule 1: If the number of values in the data set is **odd**, the median is the **middle ranked value**

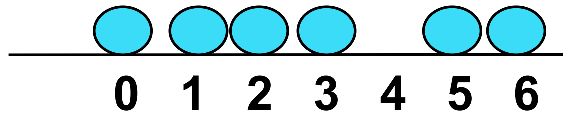
Rule 2: If the number of values in the data set is **even**, the median is the **mean** (average) of the **two middle ranked values**

Stat Joke: "You know how dumb the average person is? Well, by definition, half the population is dumber than that!"

+Mode

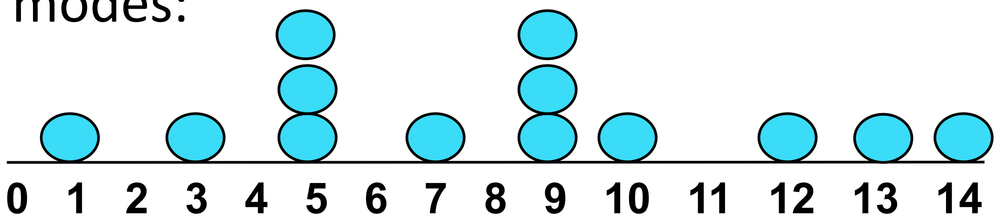
- A measure of central tendency
- Value that occurs most often (the most frequent)
- Not affected by extreme values
- Used for either numerical or categorical (nominal) data
- Unlike mean and median, there may be no unique (single) mode for a given data set

An example of no mode:



An example of several modes:

Modes = 5 and 9



Australian Bureau of Statistics

Search

Statistics | **Census** | **Complete your survey** | **About us**

ABS Home > Census > Quickstats

2016 Census QuickStats

Australia

Australia

Code 0 (AUST)

Search for a Community Profile

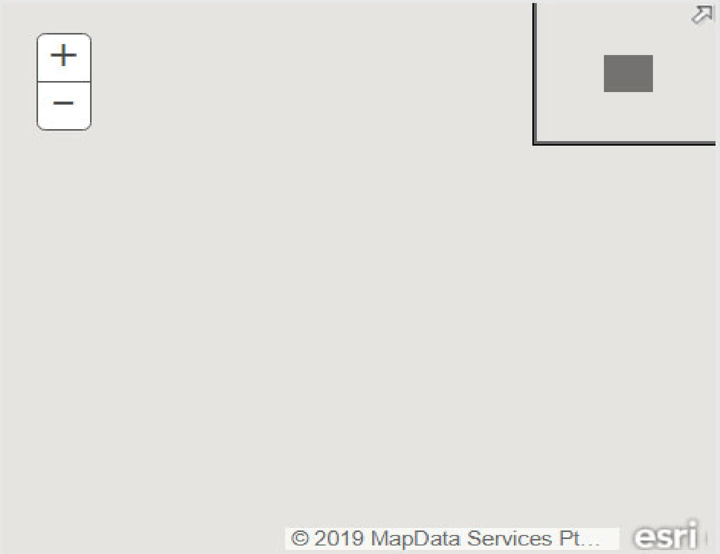
People	**23,401,892**
Male	49.3%
Female	50.7%
Median age	38

Families	**6,070,316**
Average children per family	
for families with children	1.8
for all families	0.8

All private dwellings	**9,901,496**
Average people per household	2.6
Median weekly household income	$1,438
Median monthly mortgage repayments	$1,755
Median weekly rent	$335
Average motor vehicles per dwelling	1.8

QuickStats Search

Enter a location GO



Median Sale Price	Quarterly Price Change	Median Rent	Rental Yield
n/a	n/a	n/a	n/a
Metro Melbourne $785k	Metro Melbourne -1.1%	Metro Melbourne $450	Metro Melbourne 3%

Detailed statistics below to be updated soon for the June 2019 quarter.

Sales Data

Insight (Houses & Units)	Suburb	Metro
Clearance Rate	n/a	68.3%
Days on Market	45	42
Median price by Bedrooms	**Suburb**	**Metro**
2 Bedrooms	n/a	$865,000
3 Bedrooms	n/a	$730,000
4 Bedrooms	n/a	$870,000

Rental Data

Median rent by Bedrooms	Suburb	Metro
2 Bedrooms	n/a	$495
3 Bedrooms	n/a	$425
4 Bedrooms	n/a	$480

https://reiv.com.au/market-insights/suburb/melbourne

+Quartiles (Location of data)

Similar to the median, we find a quartile by determining the value in the appropriate **position** in the **ranked** data, where:

First quartile position: $\mathbf{Q_1 = (n+1)/4}$

Second quartile position: $\mathbf{Q_2 = (n+1)/2}$ (the median)

Third quartile position: $\mathbf{Q_3 = 3(n+1)/4}$

where *n* is the number of observed values (sample size)

P^{th} Percentile = (n+1)*P/100

Q_1 = 25^{th} Percentile = (n+1)*25/100 = (n+1)/4

+Quartiles

Use the following rules to calculate the quartiles:

Rule 1 If the result is an integer, then the quartile is equal to the ranked value. For example, if the sample size is n = 7, the first quartile, Q1, is equal to the (7 + 1)/4 = 2, second-ranked value

Rule 2 If the result is a fractional half (2.5, 4.5, etc.), then the quartile is equal to the mean of the corresponding ranked values. For example, if the sample size is n = 9, the first quartile, Q1, is equal to the (9 + 1)/4 = 2.5 ranked value, halfway between the second- and the third-ranked values

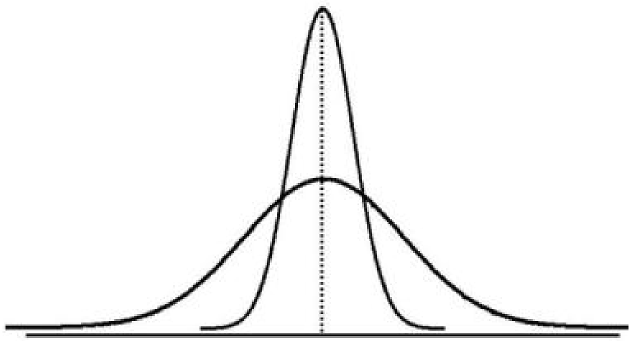
Rule 3 If the result is neither an integer nor a fractional half, round the result to the nearest integer and select that ranked value. For example, if the sample size is n = 10, the first quartile, Q1, is equal to the (10 + 1)/4 = 2.75 ranked value. Round 2.75 to 3 and use the third-ranked value

+Measures of Variation

Variation

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

Measures of variation gives information on the **spread** or **variability** of the data values



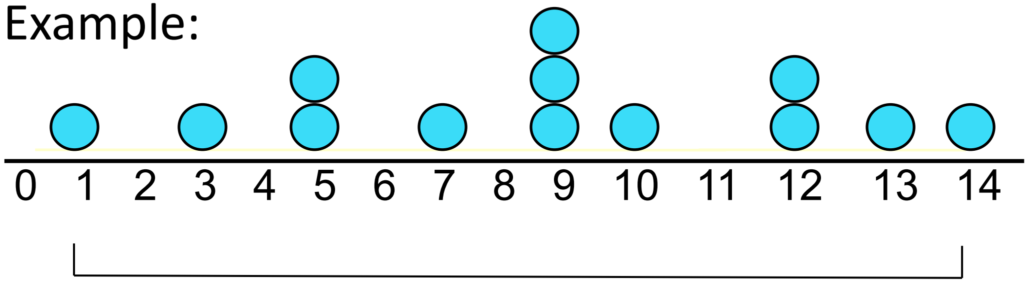
e.g. same centre, different variation

+Range

- Simplest measure of variation
- Difference between the largest and smallest values in data set
- Ignores the distribution of the data
- Like the Mean, the Range is sensitive to outliers

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

Example:



Range = 14 - 1 = 13

+Interquartile Range

Like the Median, Q_1 and Q_3, the IQR is a **resistant summary measure** (resistant to the presence of extreme values)

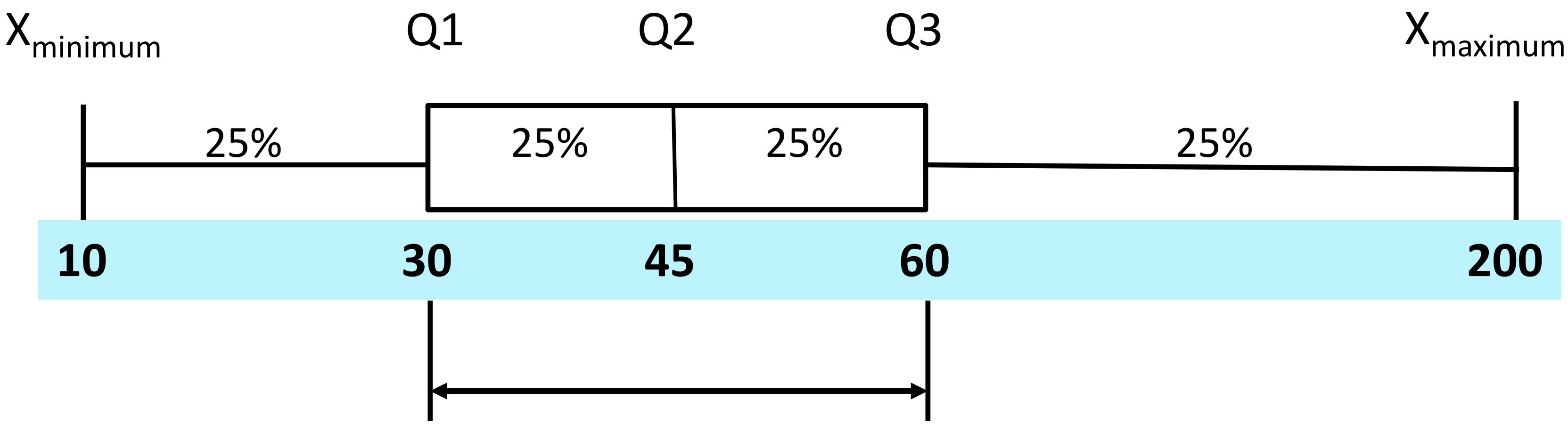
Eliminates outlier problems by using the **interquartile range**, as high- and low-valued observations are removed from calculations

IQR = 3rd quartile – 1st quartile

$$\mathbf{IQR = Q_3 - Q_1}$$

+Interquartile Range

Example: Range = 200–10 = 190 (misleading)



IQR = 60 – 30 = 30

+Variance and Standard Deviation

The **Sample Variance** – S^2

- Measures average scatter around the mean
- Units are also squared

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

Where:
$\overline{X}$ = sample mean
n = sample size
X_i = i^{th} value of the variable X

+Variance and Standard Deviation

The **Sample Standard Deviation** – S

- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

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

Where:
$\overline{X}$ = sample mean
n = sample size
X_i = i^{th} value of the variable X

+Variance and Standard Deviation

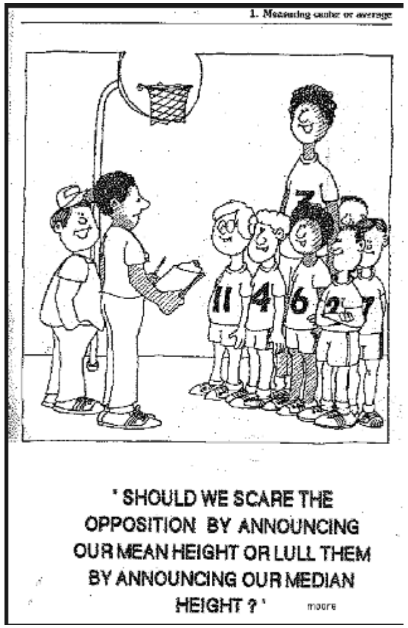
Advantages

- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight as deviations from the mean are squared

Disadvantages

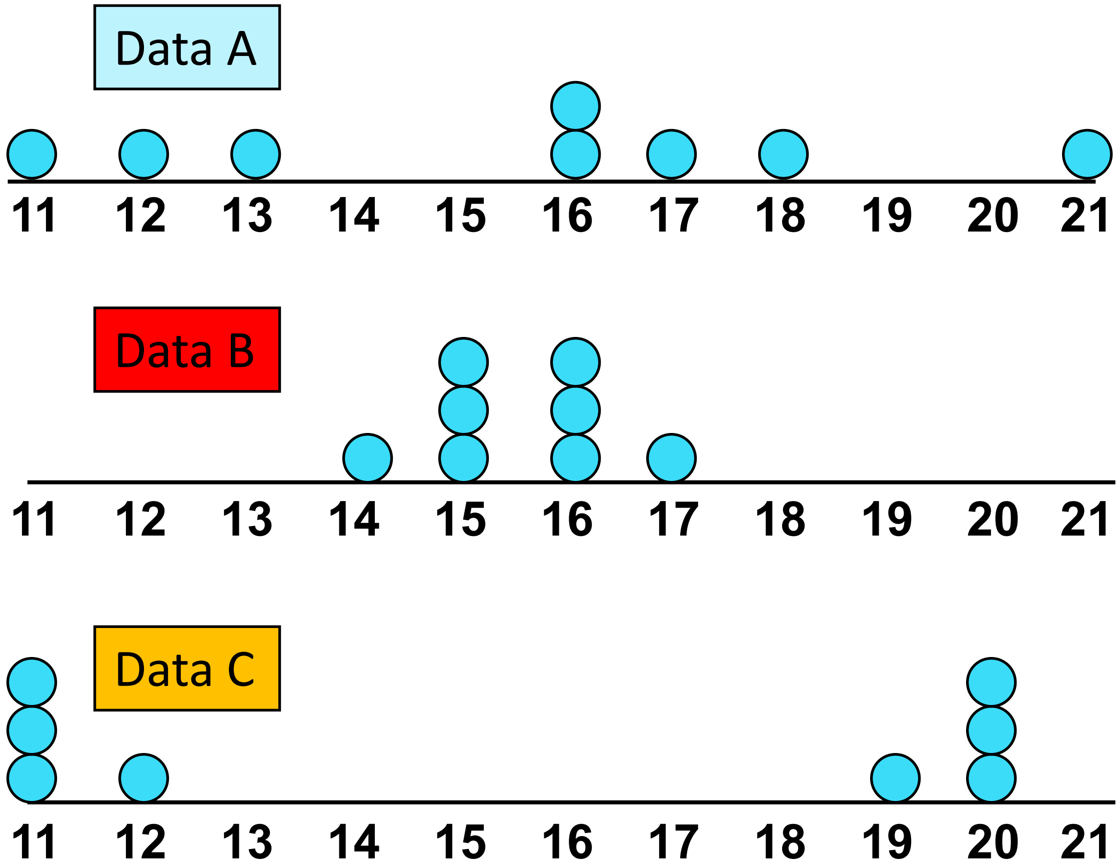
- Sensitive to extreme values (outliers)
- Measures of absolute variation not relative variation

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



Source: http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/DescriptiveStatistics/

+Comparing Standard Deviations



Mean = 15.5

S = 3.338

Mean = 15.5

S = 0.926

Mean = 15.5

S = 4.567

+Coefficient of Variation

Measures relative variation

- i.e. shows variation relative to mean

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

Always expressed as percentage (%)

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

+Shape

Mean = Median = Mode

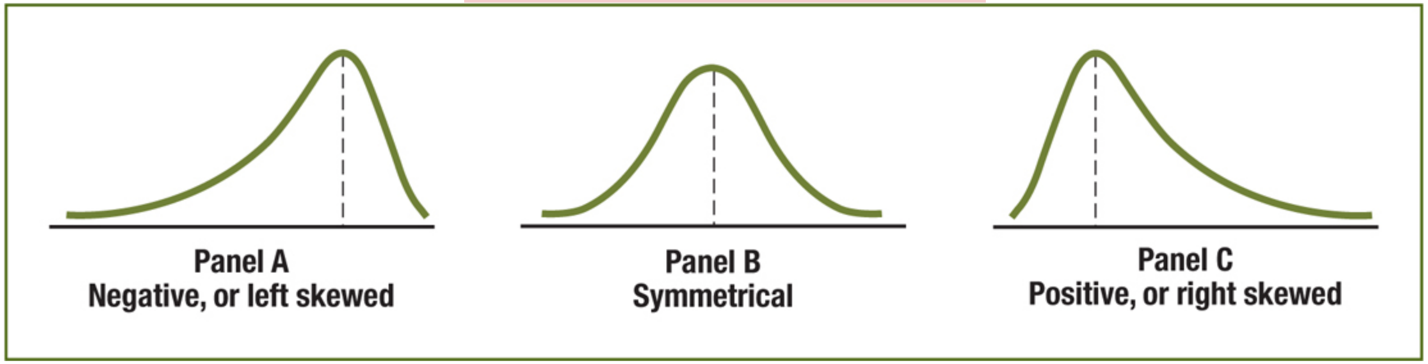


Figure 3.1
A comparison of three data sets differing in shape

Mean < Median < Mode

Mean > Median > Mode

+Shape

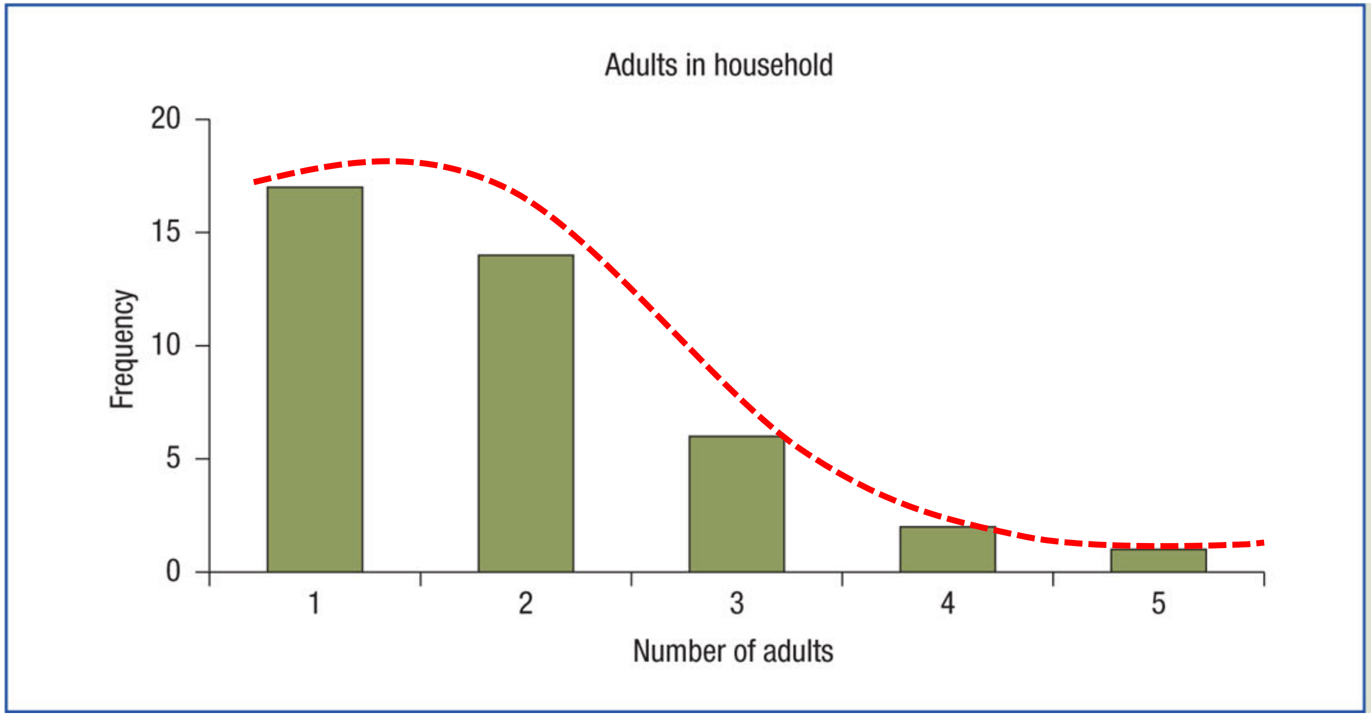


Figure 3.2
Column chart for number of adults in household

+Microsoft Excel Descriptive Statistics Output

	A	B
1	**Festival spending – international visitors**	
2		
3	Mean	743.75
4	Standard error	74.9867
5	Median	744
6	Mode	#N/A
7	Standard deviation	259.761
8	Sample variance	67476
9	Kurtosis	–1.41411
10	Skewness	–0.13236
11	Range	776
12	Minimum	343
13	Maximum	1119
14	Sum	8925
15	Count	12

Figure 3.3 Microsoft Excel summary statistics for festival expenditure

+3.2 Numerical Descriptive Measures for a Population

- Population summary measures are called parameters
- The population mean is the sum of the values in the population divided by the population size, N

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

+Population Variance and Standard Deviation

Population Variance:

- the average of the squared deviations of values from the mean

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

μ = population mean; N = population size; X_i = i^{th} value of the variable X

Population Standard Deviation:

- shows variation about the mean
- is the square root of the population variance
- has the same units as the original data

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

+Z Scores

The difference between a given observation and the mean, divided by the standard deviation

$$Z = \frac{X - \bar{X}}{S}$$

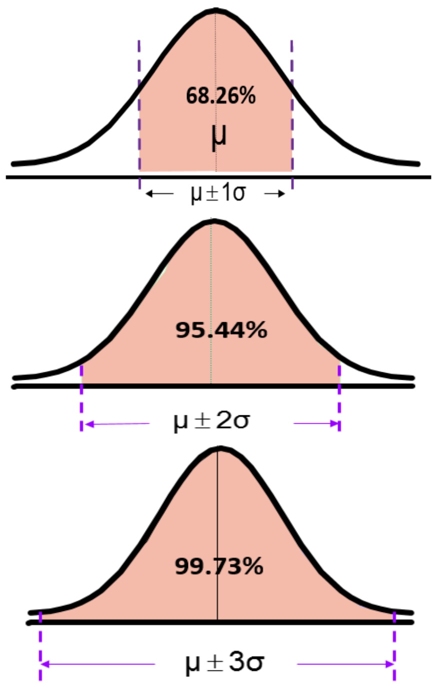
For example:

- A Z score of 2.0 means that a value is 2.0 standard deviations from the mean
- A Z score above 3.0 or below -3.0 is considered an outlier (symmetrical distribution)

+The Empirical Rule

If the data distribution is approximately bell-shaped, then the interval:

- $\mu \pm 1\sigma$ contains about 68.26% of values of the population
- $\mu \pm 2\sigma$ contains about 95.44% of values of the population
- $\mu \pm 3\sigma$ contains about 99.73% of values of the population



+The Chebyshev Rule

Interval	% of values found in intervals around the mean Chebyshev (any distribution)	Empirical rule (bell-shaped distribution)
$(\mu - \sigma, \mu + \sigma)$	At least 0%	Approximately 68%
$(\mu - 2\sigma, \mu + 2\sigma)$	At least 75%	Approximately 95%
$(\mu - 3\sigma, \mu + 3\sigma)$	At least 88.89%	Approximately 99.7%

Table 3.4
How data vary around the mean

+3.3 Calculating Numerical Descriptive Measures from a Frequency Distribution

Sometimes only a frequency distribution is available, not the raw data

Use the midpoint of a class interval to approximate the values in that class

$$\overline{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

where: n = number of values or sample size

c = number of classes in the frequency distribution

m_j = midpoint of the j^{th} class

f_j = number of values in the j^{th} class

+ 3.3 Calculating Numerical Descriptive Measures from a Frequency Distribution

Approximating the Standard Deviation

$$S = \sqrt{\frac{\sum_{j=1}^{c} (m_j - \overline{X})^2 f_j}{n-1}}$$

$$S = \sqrt{\frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1}}$$

Note: Assume that all values within each class interval are located at the midpoint of the class

3.3 Calculating Numerical Descriptive Measures from a Frequency Distribution (cont)

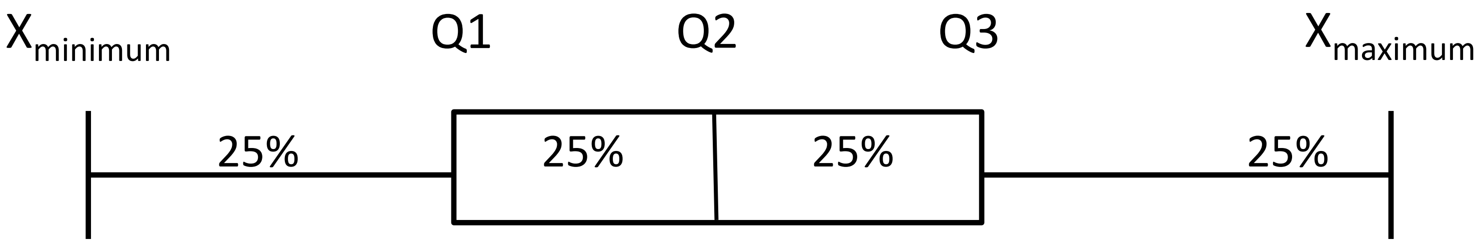
Table 3.6
Calculations needed to calculate approximations of the mean and standard deviation of the real estate prices

Asking price ($)	Frequency	Mid-point in $000s	$f_j m_j$	$f_j m_j^2$
300,000 to < 350,000	8	325	2,600	845,000
350,000 to < 400,000	17	375	6,375	2,390,625
400,000 to < 450,000	21	425	8,925	3,793,125
450,000 to < 500,000	20	475	9,500	4,512,500
500,000 to < 550,000	16	525	8,400	4,410,000
550,000 to < 600,000	6	575	3,450	1,983,750
600,000 to < 650,000	7	625	4,375	2,734,375
650,000 to < 700,000	3	675	2,025	1,366,875
700,000 to < 750,000	0	725	0	0
750,000 to < 800,000	0	775	0	0
800,000 to < 850,000	2	825	1,650	1,361,250
Totals	100		47,300	23,397,500

$$\bar{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}$$

$$S = \sqrt{\frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1}}$$

+3.4 Five-Number Summary and Box-and-Whisker Plot



Minimum($X_{smallest}$) -- Q1 -- Median -- Q3 -- Maximum ($X_{largest}$)

+Five Number Summary

Comparison	Type of distribution: Left skewed	Symmetrical	Right skewed
Distance from $X_{smallest}$ to the median versus the distance from the median to $X_{largest}$.	The distance from $X_{smallest}$ to the median is greater than the distance from the median to $X_{largest}$.	Both distances are the same.	The distance $X_{smallest}$ to the median is less than the distance from the median to $X_{largest}$.
Distance from $X_{smallest}$ to Q_1 versus the distance from Q_3 to $X_{largest}$.	The distance from $X_{smallest}$ to Q_1 is greater than the distance from Q_3 to $X_{largest}$.	Both distances are the same.	The distance from $X_{smallest}$ to Q_1 is less the distance from Q_3 to $X_{largest}$.
Distance from Q_1 to the median versus the distance from the median to Q_3.	The distance from Q_1 to the median is greater than the distance from the the median to Q_3.	Both distances are the same.	The distance from Q_1 to the median is less than the distance from the the median to Q_3.

Table 3.7 Relationships between the five-number summary and the type of distribution

+Distribution Shape and Box-and-Whisker Plots

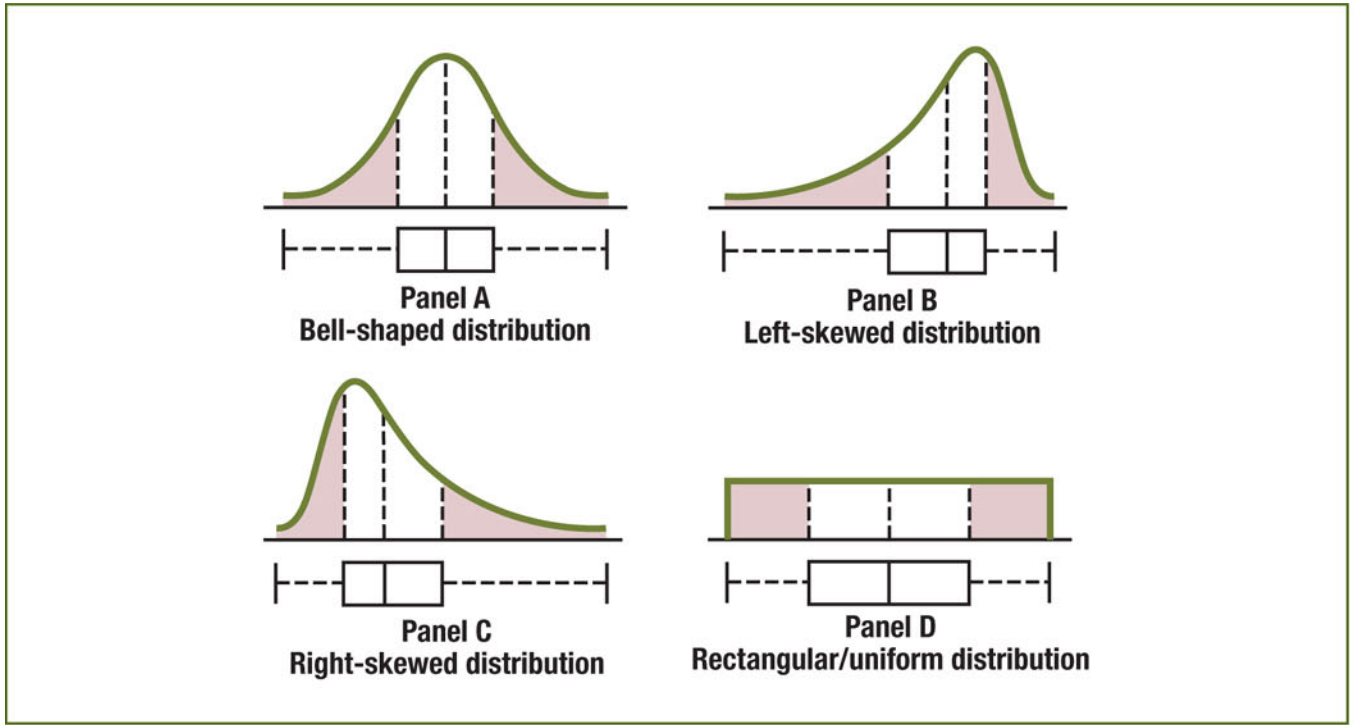


Figure 3.6
Box-and-whisker plots and corresponding polygons for four distributions

+2.4 Cross Tabulations

Table 2.11 Frequency contingency table for number of bedrooms and location

Location	Bedrooms 1	2	3	4	>4	Total
Rural	2	5	16	10	1	34
Town	4	14	29	14	5	66
Total	6	19	45	24	6	100

Table 2.12 Percentage contingency table for number of bedrooms and location based on overall total

Location	Bedrooms % 1	2	3	4	>4	Total %
Rural	2.0	5.0	16.0	10.0	1.0	34.0
Town	4.0	14.0	29.0	14.0	5.0	66.0
Total	6.0	19.0	45.0	24.0	6.0	100.0

+Side-by-Side Bar Charts

Figure 2.12
Microsoft Excel side-by-side bar chart for number of bedrooms and location

+2.4 Cross Tabulations

Location	Bedrooms 1	2	3	4	>4	Total
Rural	2	5	16	10	1	34
Town	4	14	29	14	5	66
Total	6	19	45	24	6	100

Location	Bedrooms % 1	2	3	4	>4	Total %
Rural	5.9	14.7	47.1	29.4	2.9	100.0
Town	6.1	21.2	43.9	21.2	7.6	100.0
Total	6.0	19.0	45.0	24.0	6.0	100.0

Table 2.13 Contingency table for number of bedrooms and location based on row total reported as a percentage

Location	Bedrooms % 1	2	3	4	>4	Total %
Rural	33.3	26.3	35.6	41.7	16.7	34.0
Town	66.7	73.7	64.4	58.3	83.3	66.0
Total	100.0	100.0	100.0	100.0	100.0	100.0

Table 2.14 Contingency table for number of bedrooms and location based on column total reported as a percentage

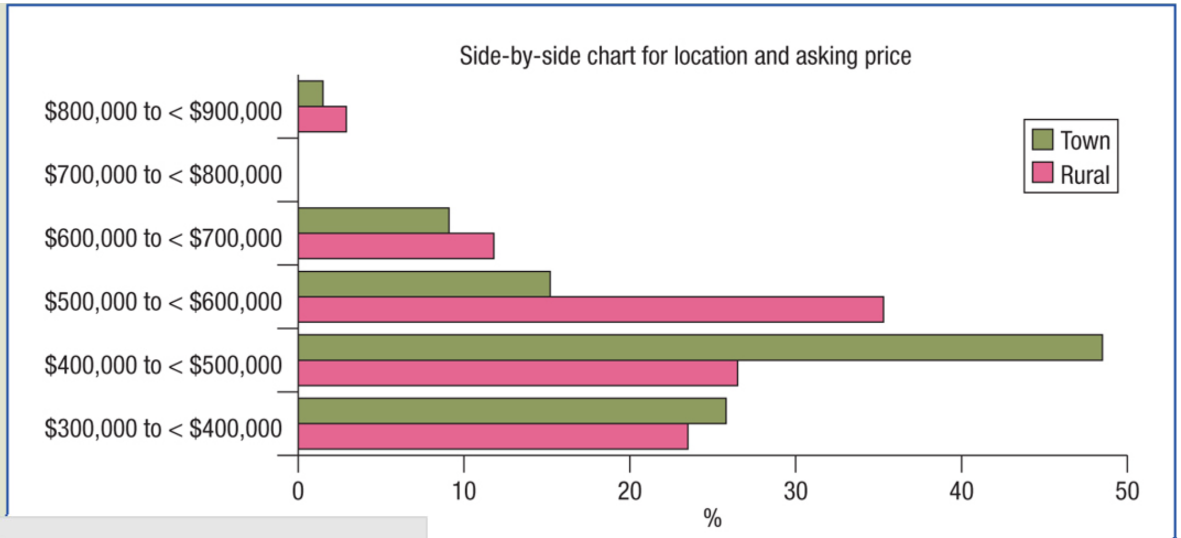
+Side-by-Side Bar Charts

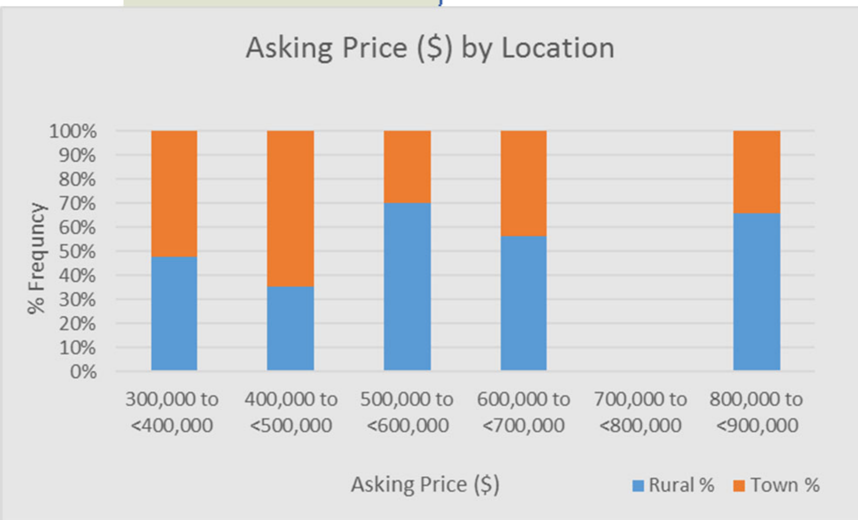
Table 2.15
Contingency table for price and location based on percentage of column total

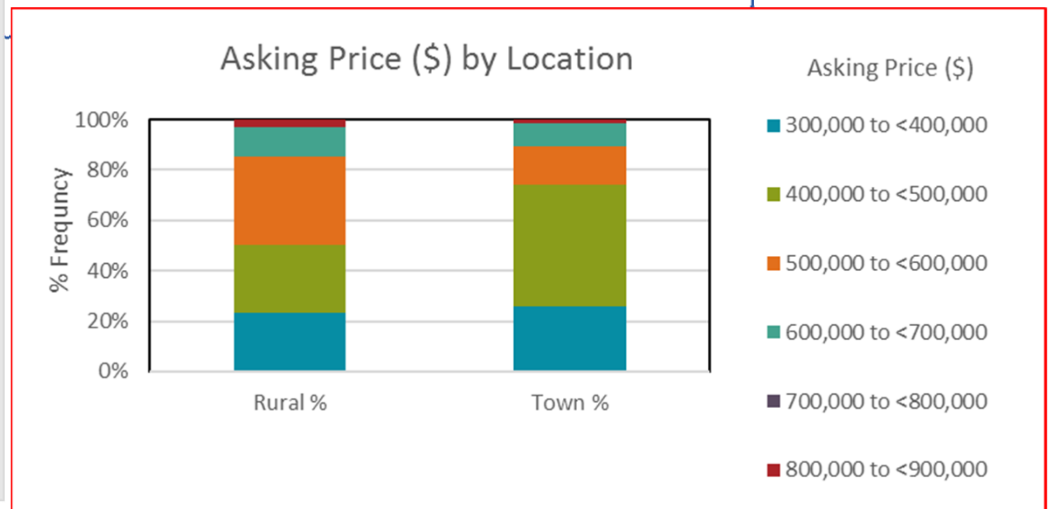
Asking price ($)	Frequency Rural	Frequency Town	Column percentage Rural	Column percentage Town
300,000 to < 400,000	8	17	23.5	25.8
400,000 to < 500,000	9	32	26.5	48.5
500,000 to < 600,000	12	10	35.3	15.1
600,000 to < 700,000	4	6	11.8	9.1
700,000 to < 800,000	0	0	0.0	0.0
800,000 to < 900,000	1	1	2.9	1.5
Total	34	66	100.0	100.0

+ Side-by-Side Bar Charts

Figure 2.13
Side-by-side chart for location and price







+2.5 Scatter Diagrams and Time-Series Plots

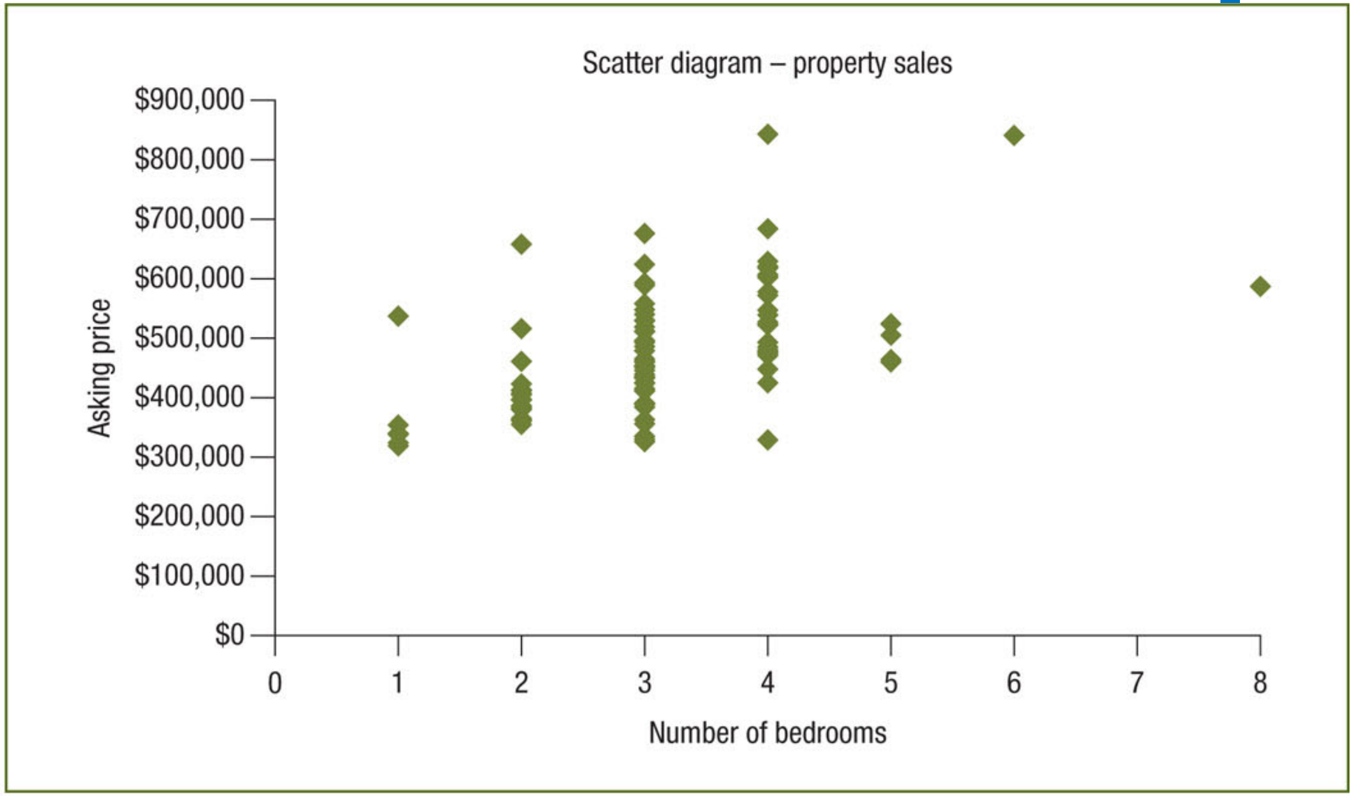
Scatter diagrams are used to examine possible relationships between two numerical variables

In a scatter diagram:

- one variable is measured on the vertical axis (Y)
- the other variable is measured on the horizontal axis (X)

+Scatter Diagrams

Figure 2.14
Microsoft Excel scatter diagram for number of bedrooms and asking price



+Time-Series Plots

A time-series plot is used to study patterns in the values of a variable over time

In a time-series plot:

- one variable is measured on the vertical axis
- the time period is measured on the horizontal axis

+Time-Series Plots

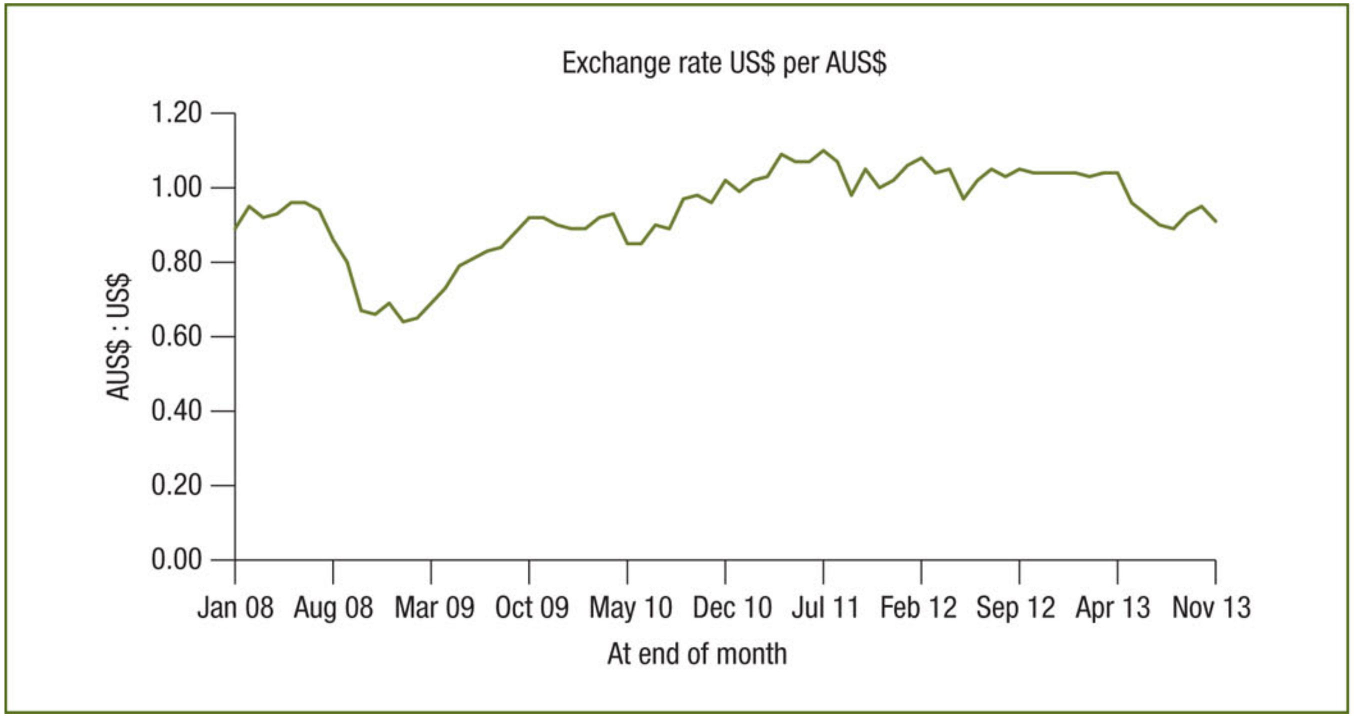


Figure 2.15
Microsoft Excel time-series plot of exchange rates: Australian dollar against US dollar 2008 to 2013

Source: Data based on *Reserve Bank of Australia, Statistics, Exchange Rates* <www.rba.gov.au> accessed December 2013.

+3.5 Covariance

The covariance is a measure of the strength and direction of the linear relationship between two numerical variables (X and Y):

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

As a covariance can have any value, it is difficult to use it as a measure of the relative strength of a linear relationship

A better, and related, measure of the relative strength of a linear relationship is the Coefficient of Correlation, *r*

3.5 Coefficient of Correlation - Calculation

The sample coefficient of correlation is the sample covariance divided by the sample deviations of *X* and *Y*

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

where:

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

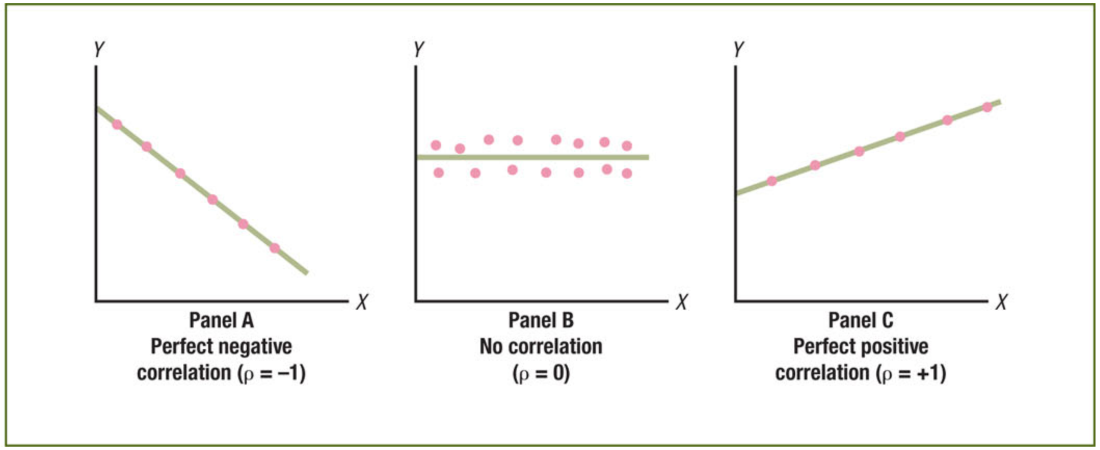
$$S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}} \qquad S_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n - 1}}$$

+3.5 Coefficient of Correlation

The coefficient of correlation measures the relative strength of a linear relationship between two numerical variables (X and Y)

Values range from -1 (perfect negative) to +1 (perfect positive)

Figure 3.7
Types of association between variables



+3.5 Coefficient of Correlation (cont)

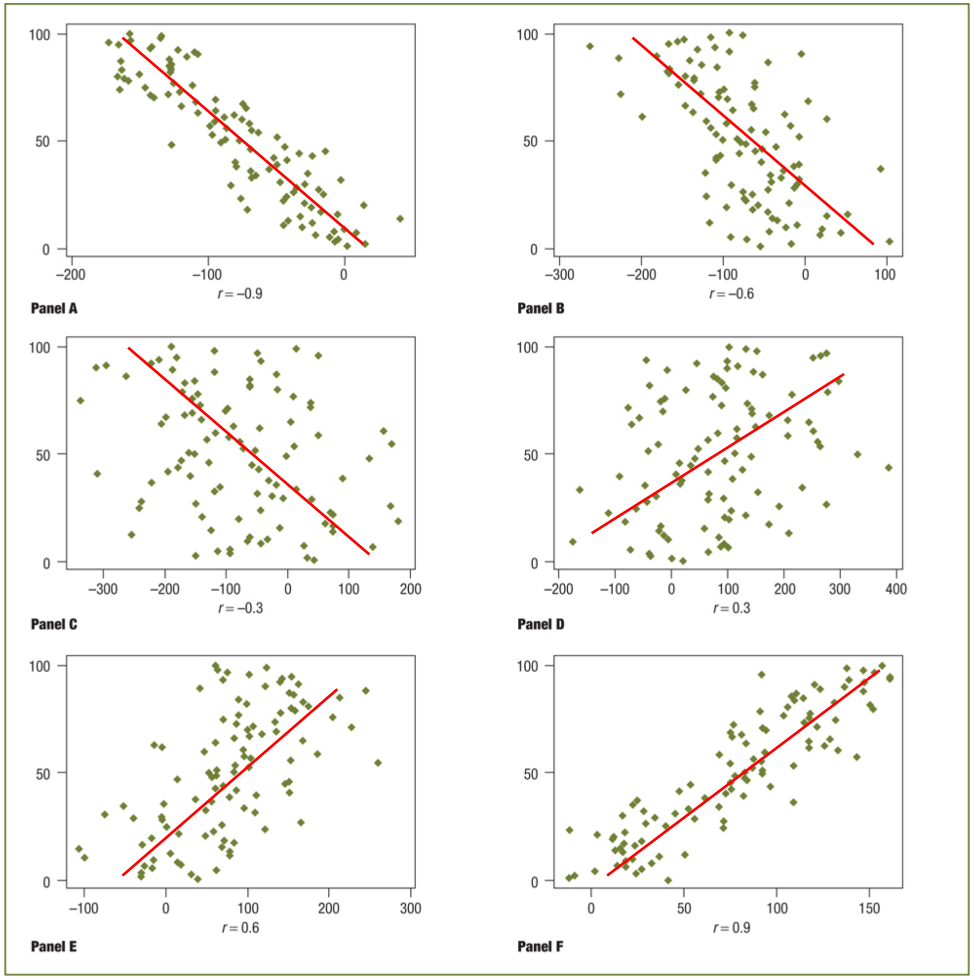


Figure 3.8 Six scatter diagrams and their sample coefficients of correlation, r

+3.6 Pitfalls in Numerical Descriptive Measures and Ethical Issues

Data analysis is *objective*

- Should report the summary measures that best meet the assumptions about the data set

Data interpretation is *subjective*

- Should document both good and bad results
- Results should be presented in a fair, objective, transparent and neutral manner
- Should not use inappropriate summary measures to distort facts
- Do not fail to report pertinent findings even if such findings do not support original argument

Nation of gamblers

Australian and New Zealand gamblers are the worst in the world, betting more money online than those of any other country...
— *The Sunday Telegraph, 22nd March, 2009*

TTD Week 3

By the end of the week make sure you...

- Understand the different summary measures and their purpose, when they can/can't be used, how to calculate, and how to interpret them.
- Read chapters 2 and 3 of the text
- Complete the suggested exercises from the text
- Summarise the key terms introduced this week

Changes to MIS770 Lecture Schedule

Topic 4 – Probability and Discrete Probability Distributions

Date: 03/12/2018 (Tuesday, Week 4)

Time: 6-7:50pm

Venue – LT13 (HC2.005)

No MIS770 lecture on 04/12/2019

Note: This is change is just for Week 4

Question

You have collected data on the number of complaints for Categorical
six different brands of automobiles sold in Australia in 2007 Categorical
and in 2017. Which of the following is the best for presenting the data?

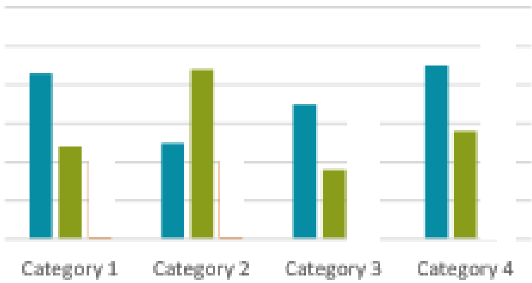
A) A time-series plot.

B) A stem-and-leaf display.

C) A contingency table.

D) A side-by-side bar chart.

Year	Brand 1	Brand 2	Brand 3	Brand 4		
2007						
2017						
Total						



+Question

Which of the following is sensitive to extreme values?

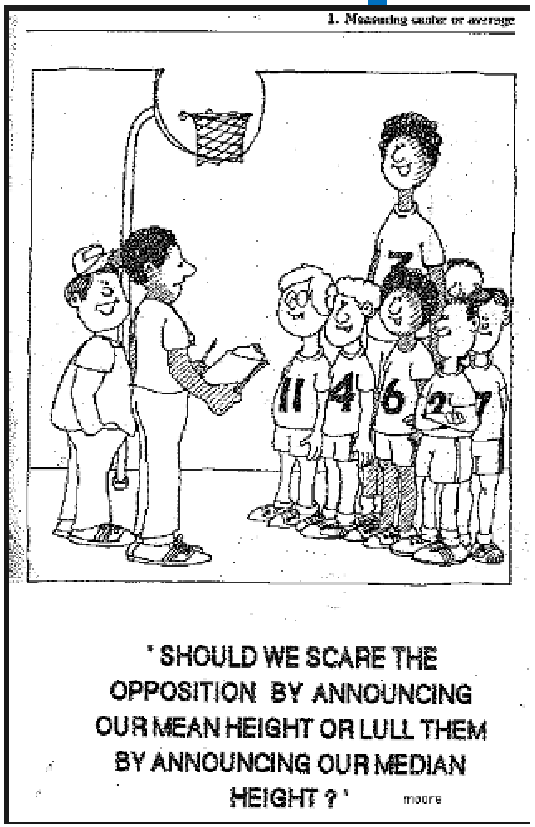
A) The median.

B) The arithmetic mean.

C) The interquartile range.

D) The 1st quartile.

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} :$$



Source: http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/DescriptiveStatistics/

Question

Numerical

You have collected data on the approximate retail price (in $) and the energy cost per year (in $) of 15 refrigerators. Which of the following is the best for presenting the data?

Numerical

A) A contingency table.

B) A scatter plot.

C) A side-by-side bar chart.

D) A pie chart.

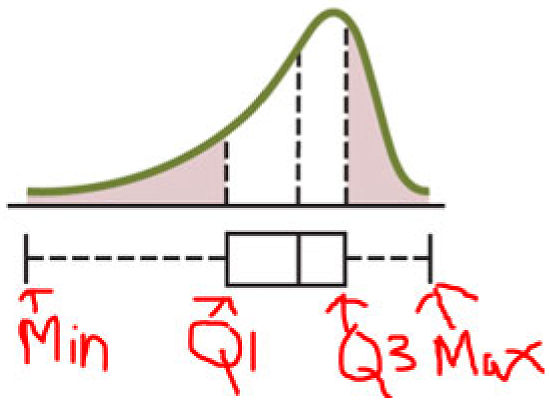
Question

True or False:

In left-skewed distributions, the distance from the smallest observation to *Q1* exceeds the distance from *Q3* to the largest observation.

A) True

B) False



Question

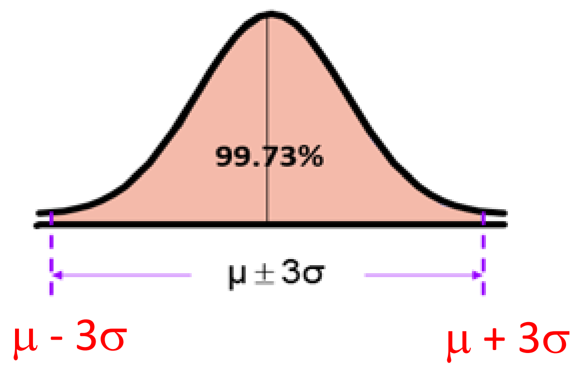
According to the empirical rule, if the data form a "bell-shaped" distribution, ________ percent of the observations will be contained within 3 standard deviations around the arithmetic mean.

A) 68.26

B) 75.00

C) 99.7

D) 95.0



Question

Which of the following is NOT sensitive to extreme values?

A) The interquartile range.

B) The standard deviation.

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

C) The coefficient of variation.

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

D) The range.

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