Lecture 2: Describing Data

Sravani Vadlamani

QMB 3200

08/27/2020

Review of Chapter 1

- Levels of Measurement
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Types of Variables
  - Numerical
    - Discrete vs Continuous
  - Categorical
    - Nominal vs Ordinal

Review of Chapter 1

- Population vs Sample
- Sampling Strategies
  - Simple Random Sampling
  - Convenience Sampling
  - Stratified Sampling
  - Cluster Sampling
- Observational Studies vs Experiments
  - Confounding Variables

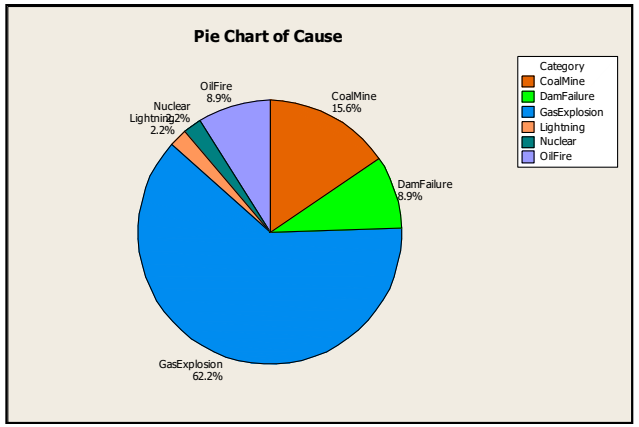
Data Display

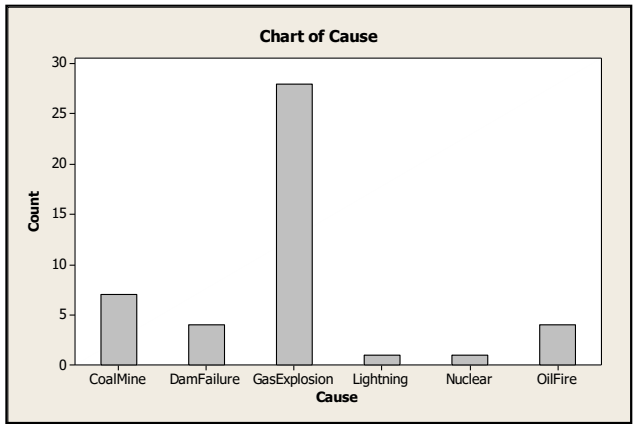
Suggested Reading

- Chapter 2- From the Textbook

Descriptive Statistics

- **Methods** of organizing, summarizing, and presenting data in a convenient and informative way. These methods include:
  - Graphical techniques (Pie charts, Histograms, etc.)





  - Numerical techniques (Mean, Standard deviation, etc.)

Graphical Methods for Describing Quantitative Data

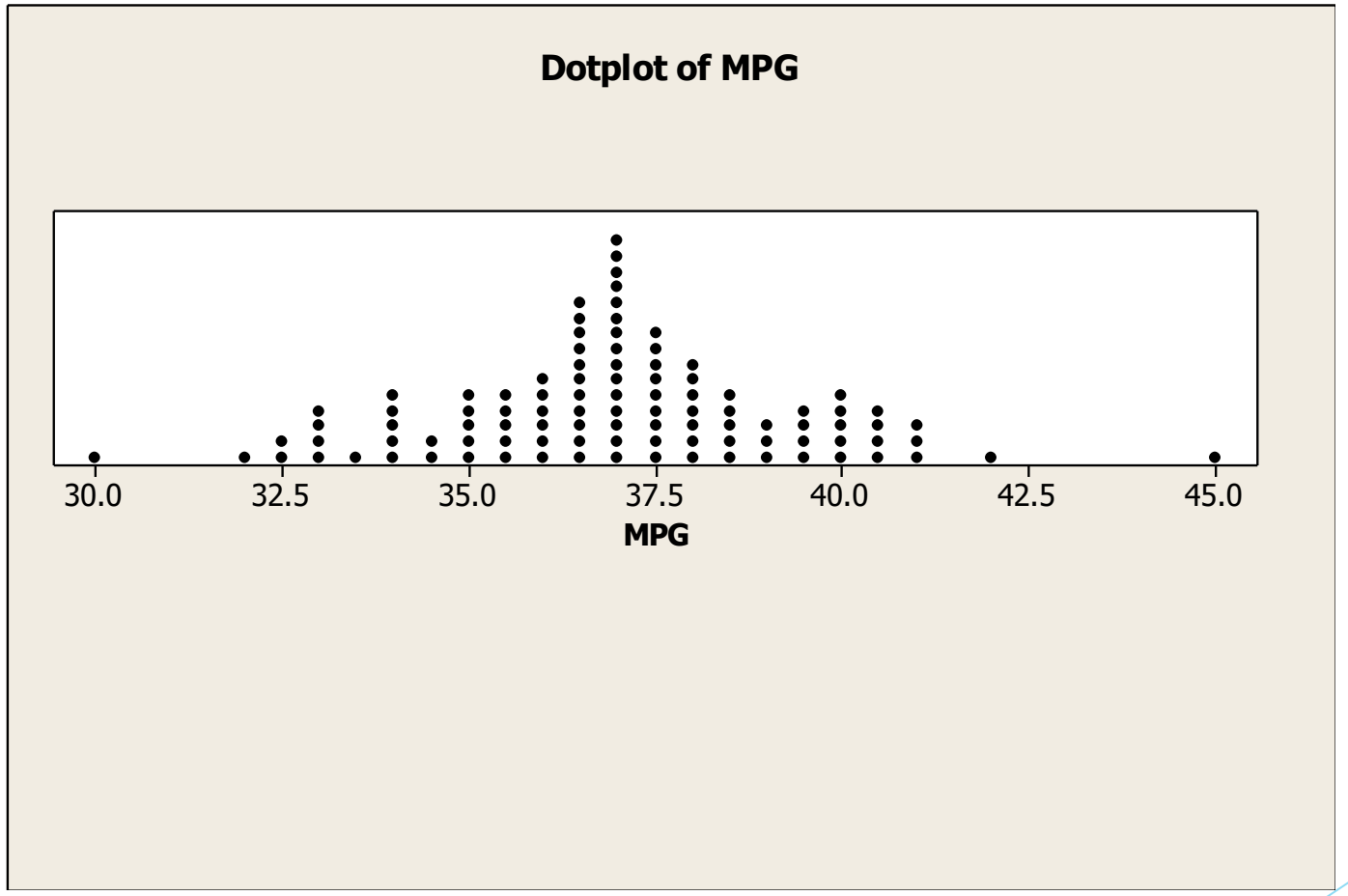
- Dot plots
- Histograms

Example

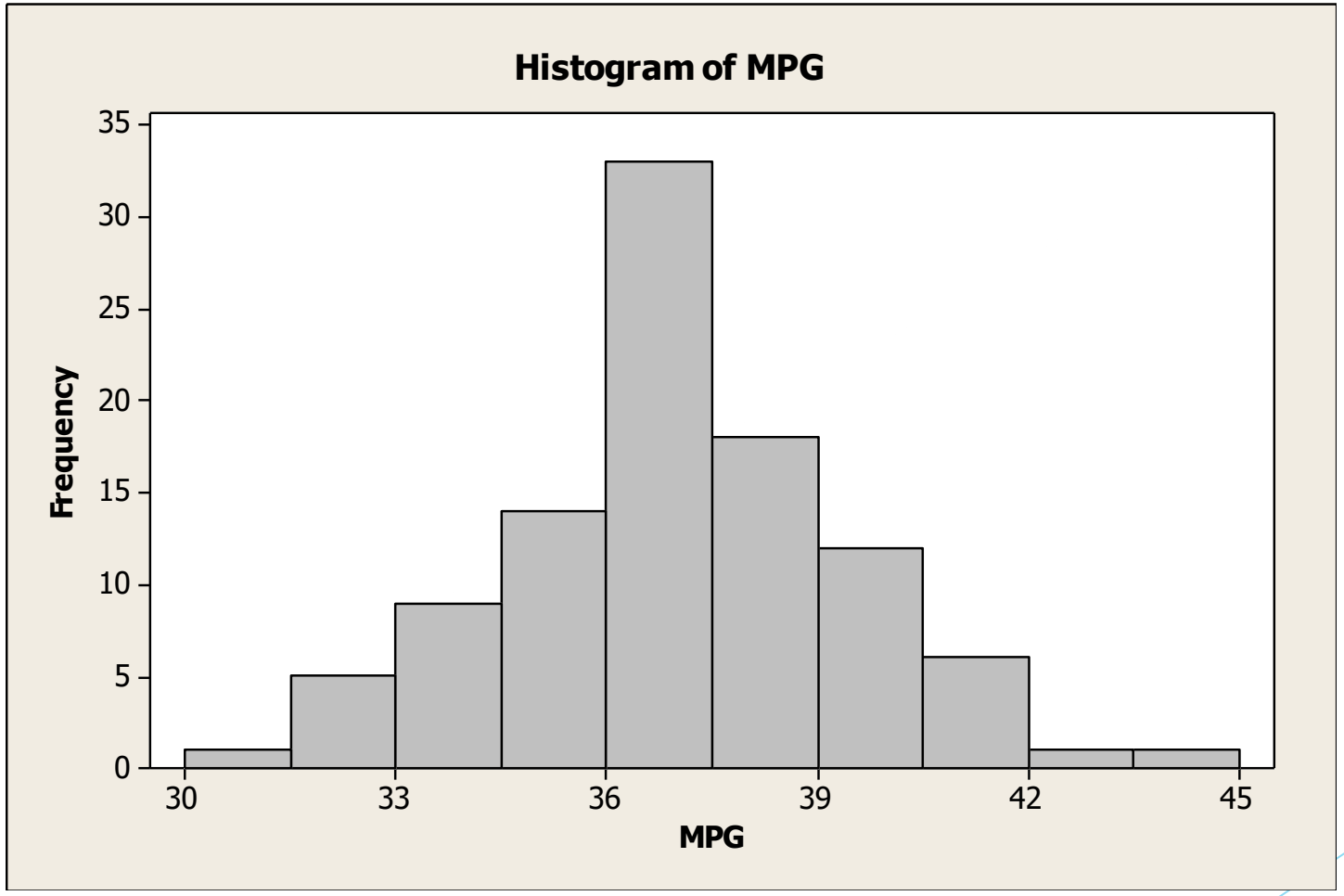
TABLE 2.2 **EPA Mileage Ratings on 100 Cars**

36.3	41.0	36.9	37.1	44.9	36.8	30.0	37.2	42.1	36.7
32.7	37.3	41.2	36.6	32.9	36.5	33.2	37.4	37.5	33.6
40.5	36.5	37.6	33.9	40.2	36.4	37.7	37.7	40.0	34.2
36.2	37.9	36.0	37.9	35.9	38.2	38.3	35.7	35.6	35.1
38.5	39.0	35.5	34.8	38.6	39.4	35.3	34.4	38.8	39.7
36.3	36.8	32.5	36.4	40.5	36.6	36.1	38.2	38.4	39.3
41.0	31.8	37.3	33.1	37.0	37.6	37.0	38.7	39.0	35.8
37.0	37.2	40.7	37.4	37.1	37.8	35.9	35.6	36.7	34.5
37.1	40.3	36.7	37.0	33.9	40.1	38.0	35.2	34.8	39.5
39.9	36.9	32.9	33.8	39.8	34.0	36.8	35.0	38.1	36.9

Dot plots

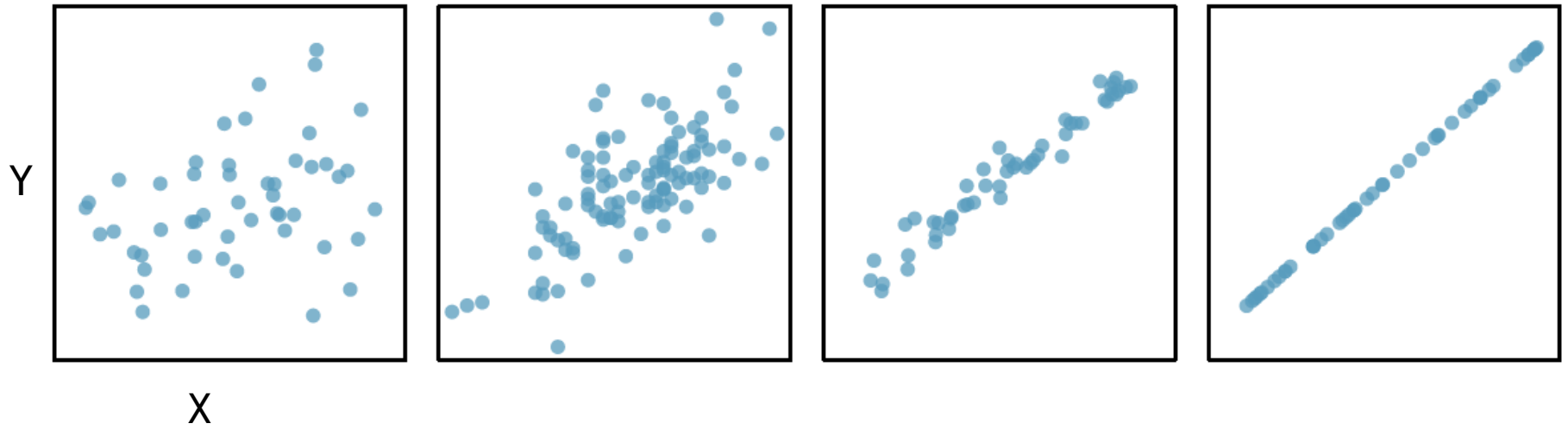


Histograms


Histogram of MPG
Frequency
35
30
25
20
15
10
5
0
30
33
36
39
42
45
MPG

Scatter Plots

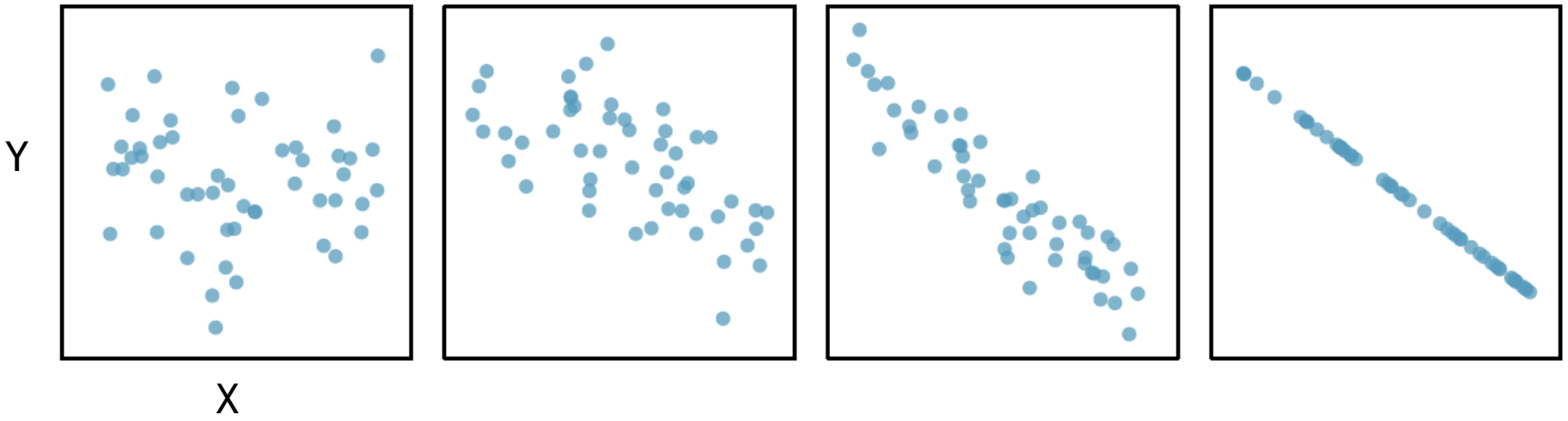
Examine relation between two variables



X = Total Income
Y = Loan Amount

Scatter Plots

Examine relation between two variables



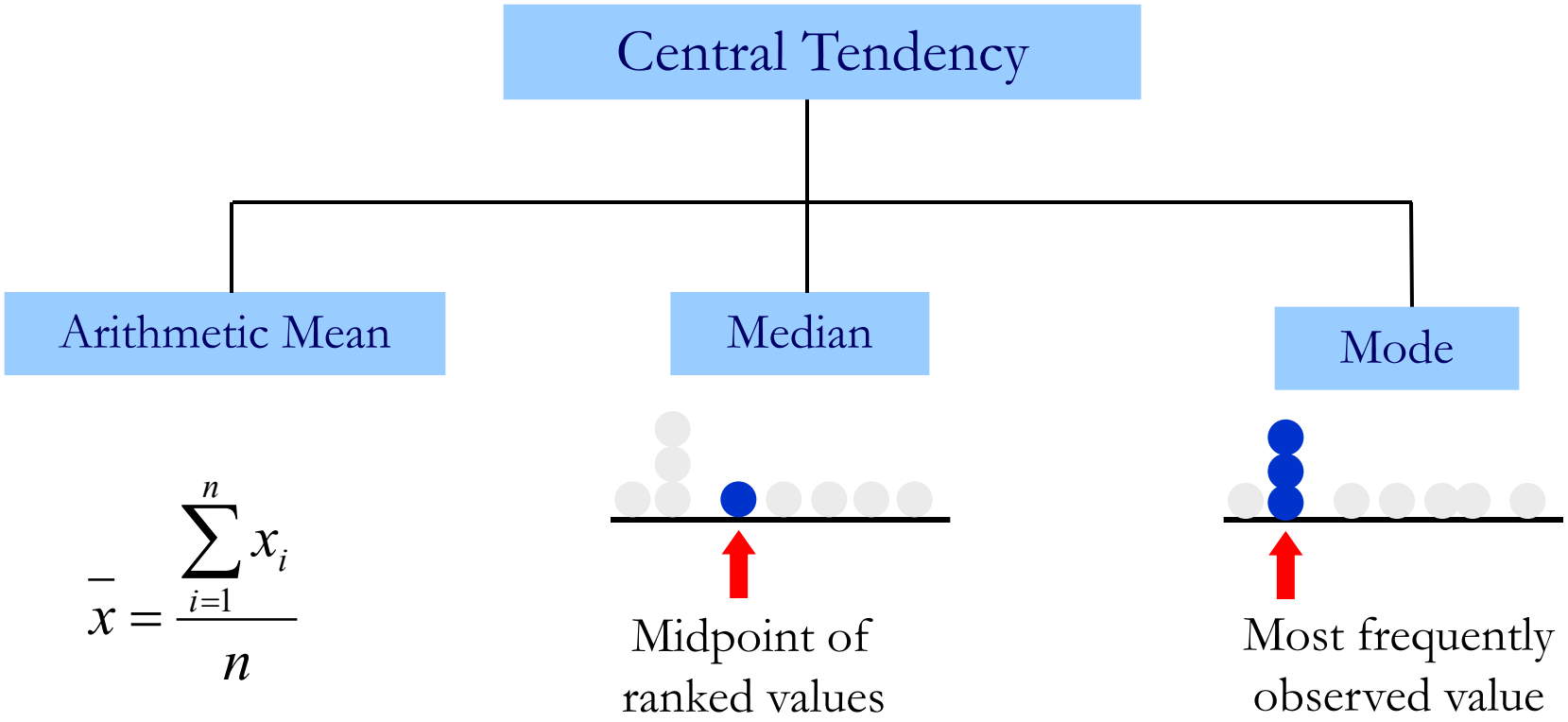
X = Total Debt
Y = Loan Interest

Numerical Methods for Describing Quantitative Data

The measures are those help to:

- Locate the *"center"* of the relative frequency distribution
  - (measures of central tendency)
- Measure *"spread"* around the center
  - (measures of variation)
- Describe the *"relative position"* of an observation
  - (measures of relative standing)

Measures of Central Tendency



Mean

	Population	Sample
Size	N	n
Mean	μ	$\bar{x}$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

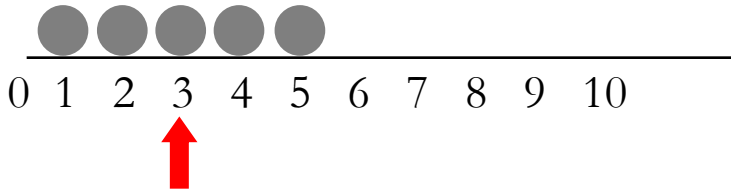
Population Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Sample Mean

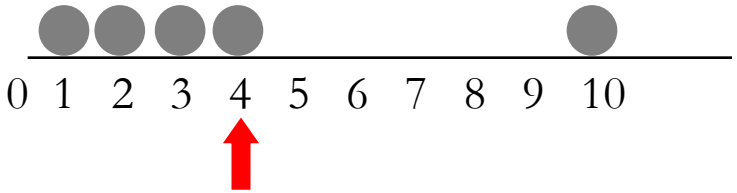
Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$



Mean = 3

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

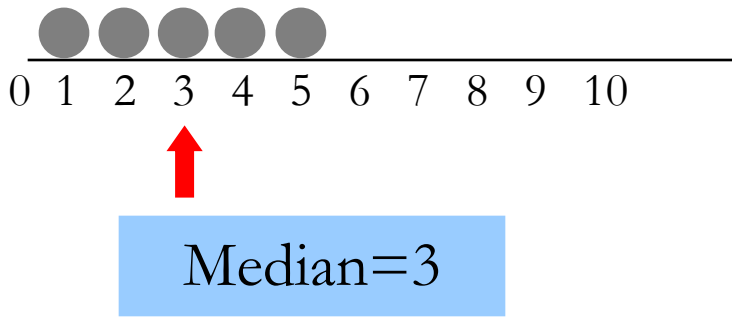


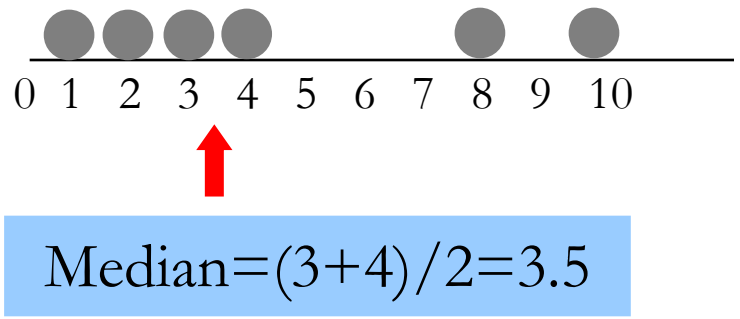
Mean = 4

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

Median

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



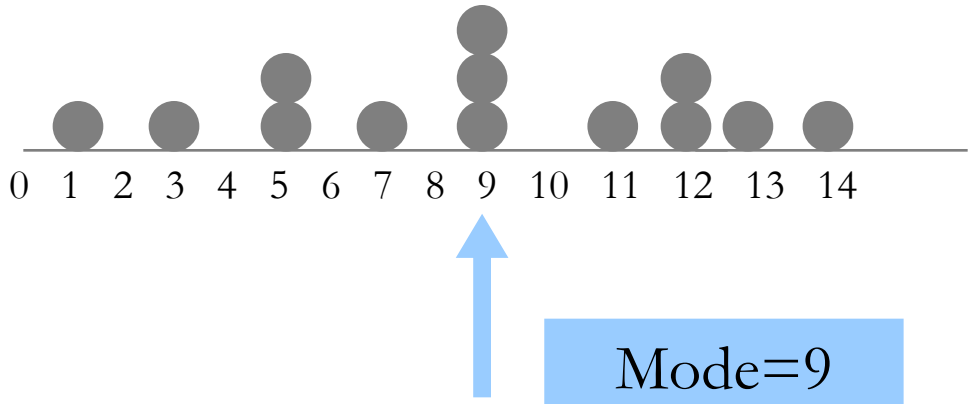


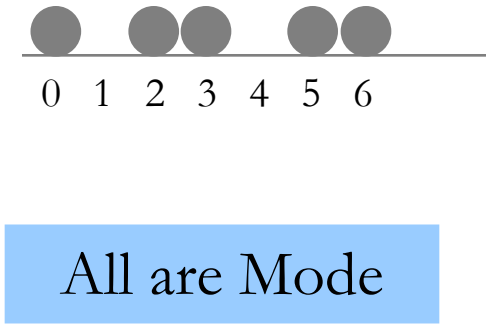
If the number of values is odd, the median is the middle number.

If the number of values is even, the median is the average of the two middle numbers.

Mode

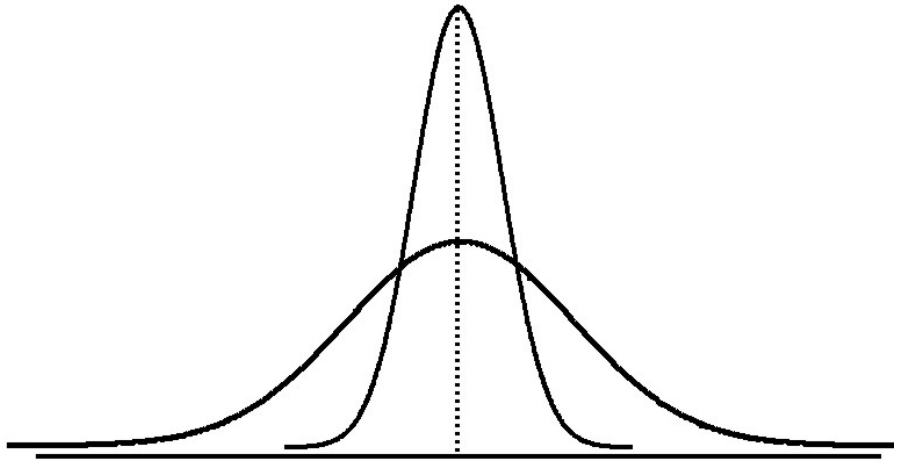
- Value that occurs most often





Measures of Variation

- Measures of variation give information on the **spread** or **variability** of the data values.
  - Range
  - Standard deviation
  - Variance

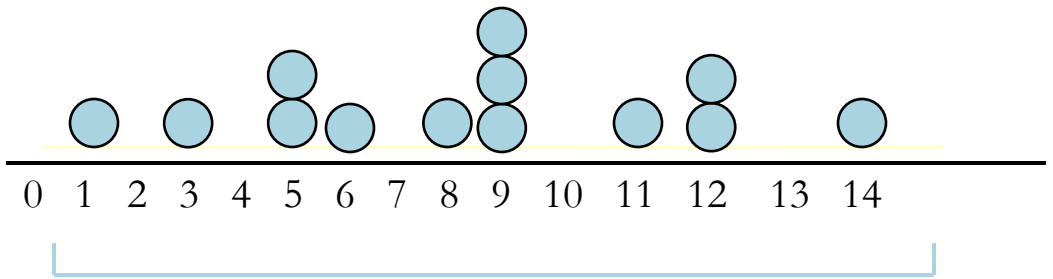


Same center, different variation

Range

$$\text{Range} = x_{\text{Largest}} - x_{\text{Smallest}}$$

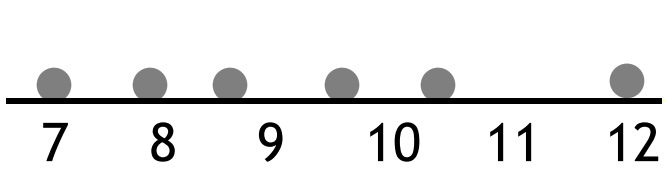
Example:



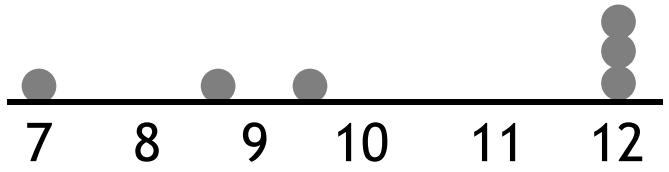
Range = 14 - 1 = 13

Disadvantages of the Range

- 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

Variance – Standard deviation

- Average of squared deviations of values from the mean
  - Sample variance and sample standard deviation

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Example: Sample Standard Deviation

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

$n = 8$ $\quad$ Mean $= \bar{x} = 16$

$$S = \sqrt{\frac{(10 - \bar{X})^2 + (12 - \bar{X})^2 + (14 - \bar{X})^2 + \ldots + (24 - \bar{X})^2}{n - 1}}$$

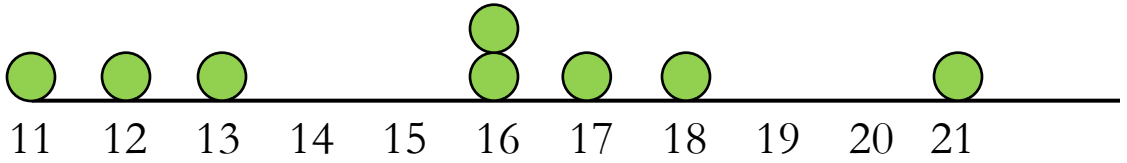
$$= \sqrt{\frac{(10 - 16)^2 + (12 - 16)^2 + (14 - 16)^2 + \ldots + (24 - 16)^2}{8 - 1}}$$

$$= \sqrt{\frac{130}{7}} = 4.3095$$

A measure of the "average" scatter around the mean

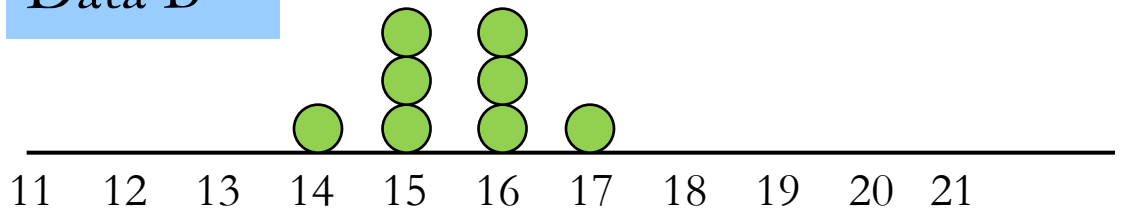
Comparing Standard Deviations

Data A



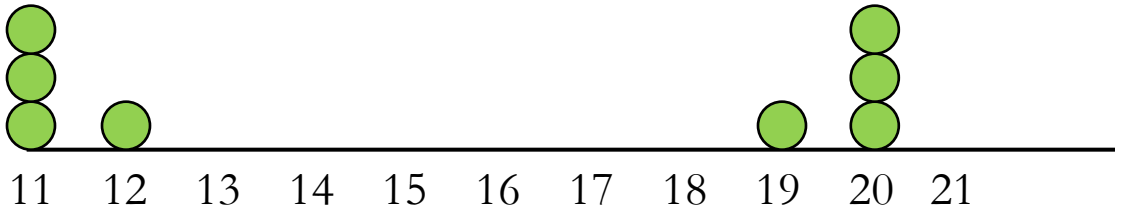
Mean = 15.5
S = 3.338

Data B



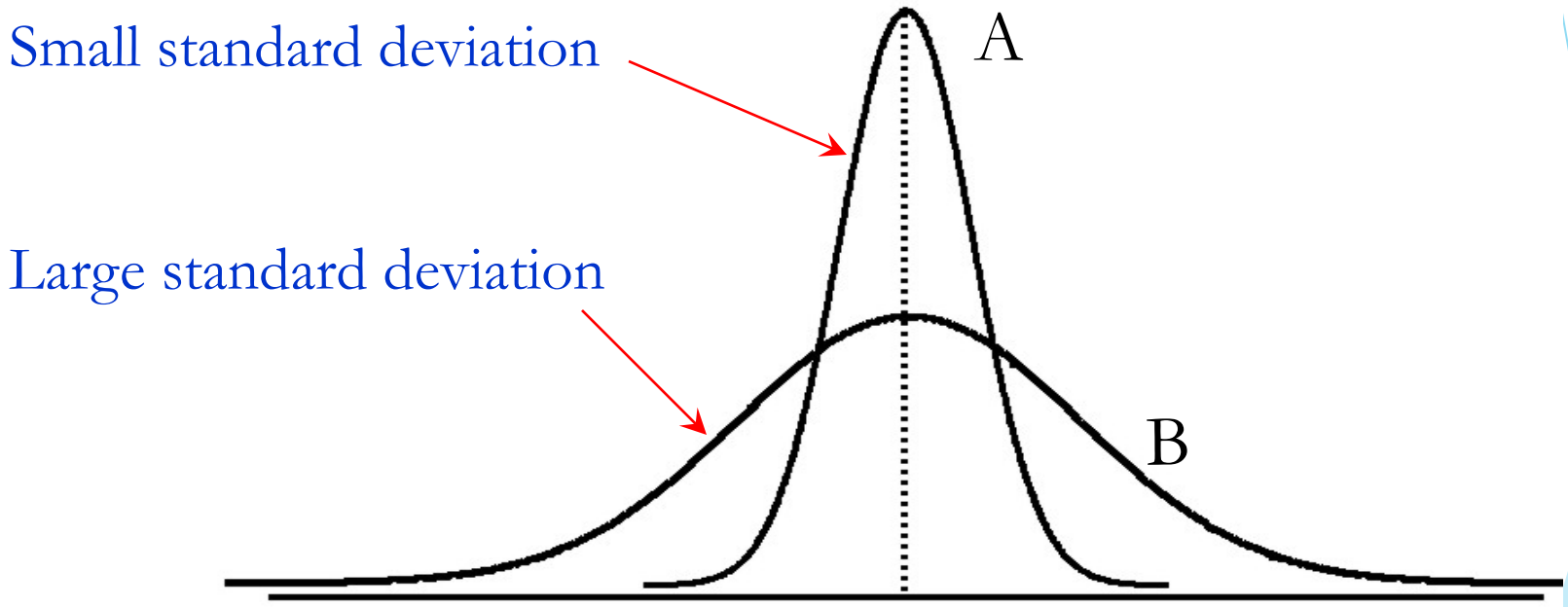
Mean = 15.5
S = 0.926

Data C



Mean = 15.5
S = 4.567

Measuring variation



Example1 – Calculate Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
4			
4			
5			
5			
5			
5			
5			
6			
6			
7			
Sum = 52	Sum = 52	Sum =	Sum =

Example – Calculate Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
4	5.2		
4	5.2		
5	5.2		
5	5.2		
5	5.2		
5	5.2		
5	5.2		
6	5.2		
6	5.2		
7	5.2		
Sum = 52	Sum = 52	Sum =	Sum =

Example – Calculate Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
4	5.2	-1.2	
4	5.2	-1.2	
5	5.2	-0.2	
5	5.2	-0.2	
5	5.2	-0.2	
5	5.2	-0.2	
5	5.2	-0.2	
6	5.2	0.8	
6	5.2	0.8	
7	5.2	1.8	
Sum = 52	Sum = 52	Sum =	Sum =

Example – Calculate Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
4	5.2	-1.2	1.44
4	5.2	-1.2	1.44
5	5.2	-0.2	.04
5	5.2	-0.2	.04
5	5.2	-0.2	.04
5	5.2	-0.2	.04
5	5.2	-0.2	.04
6	5.2	0.8	.64
6	5.2	0.8	.64
7	5.2	1.8	3.24
Sum = 52	Sum = 52	Sum = 0	Sum =7.6

Example Solution

Variance

$$s^2 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1} = \frac{7.6}{9} = .84$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{.84} = .92$$

Example2

Data

2,2,3,3,4,5,5,5,5,6

Calculate mean, median, mode, variance and standard deviation

Example2 – Calculate Mean, Median, Mode, Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
2			
2			
3			
3			
4			
5			
5			
5			
5			
6			
Sum = 40	Sum =	Sum =	Sum =

Example2 – central tendency solution

Data

2,2,3,3,4,5,5,5,5,6

Mode = 5

Median = 4.5

Mean = 4

Example2 –Variance and Standard Deviation

X_i	$\overline{X}$	$\left(X_i - \overline{X}\right)$	$\left(X_i - \overline{X}\right)^2$
2	4	-2	4
2	4	-2	4
3	4	-1	1
3	4	-1	1
4	4	0	0
5	4	1	1
5	4	1	1
5	4	1	1
5	4	1	1
6	4	2	4
Sum = 40	Sum = 40	Sum =0	Sum =18

Example2 Variability Solution

Variance

$$s^2 = \frac{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}{N-1} = \frac{18}{9} = 2$$

Standard Deviation

$$s = \sqrt{s^2} = \sqrt{2} = 1.41$$

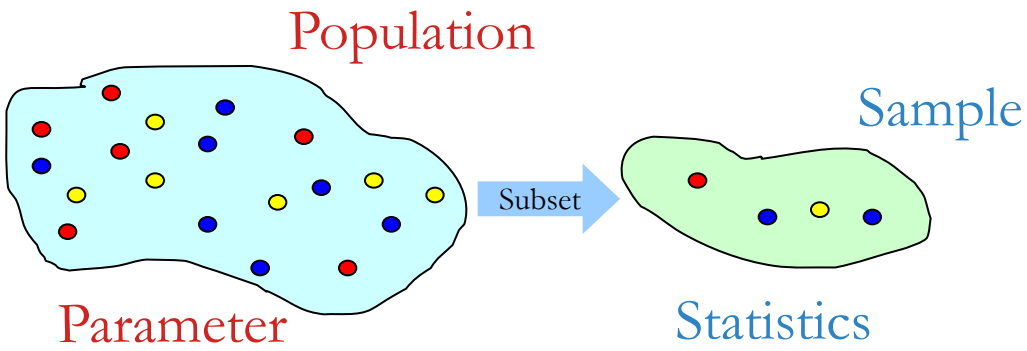
Example

Find the mean, median, mode, range and standard deviation for the following list of values:

13, 18, 13, 14, 13, 16, 14, 21, 13

Mean	
Median	
Mode	
Range	
Standard Deviation	

Population vs Sample



$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

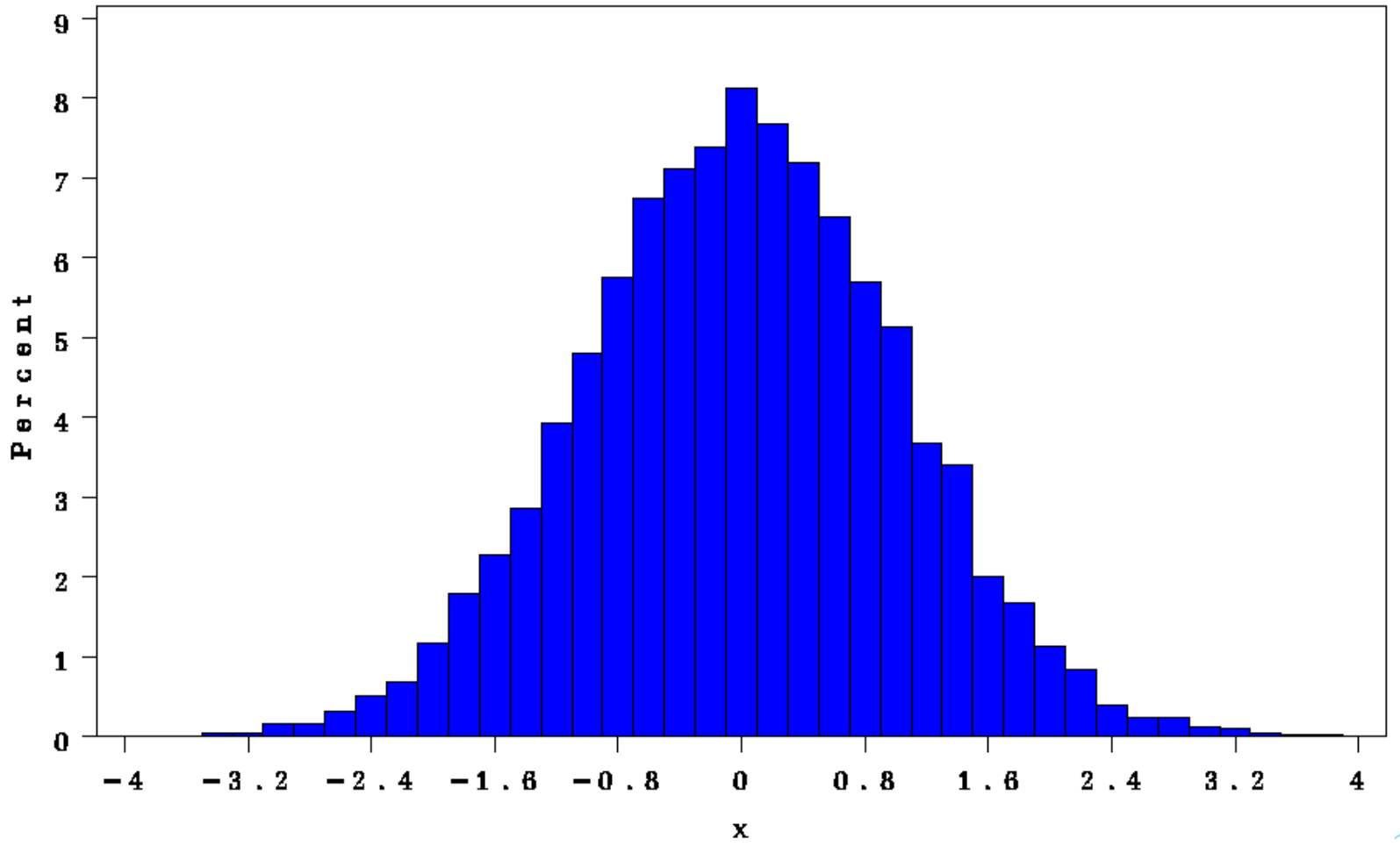
Describing Distributions

- Terms
  - Symmetric
    - Same shape on both sides of the center
  - Modality - # of peaks
    - Unimodal – one peak
    - Bimodal – two peaks
    - Rectangular Distribution – no peaks

Describing Distributions

- Terms
  - Skewness – symmetry
    - Negative – fewer scores to the left, skew < 0
    - Positive – fewer scores to the right, skew > 0
  - Kurtosis – concentration of scores
    - Mesokurtic – normal, kurtosis = 0
    - Platykurtic – flatter, kurtosis < 0
    - Leptokurtic – more peaked, kurtosis > 0

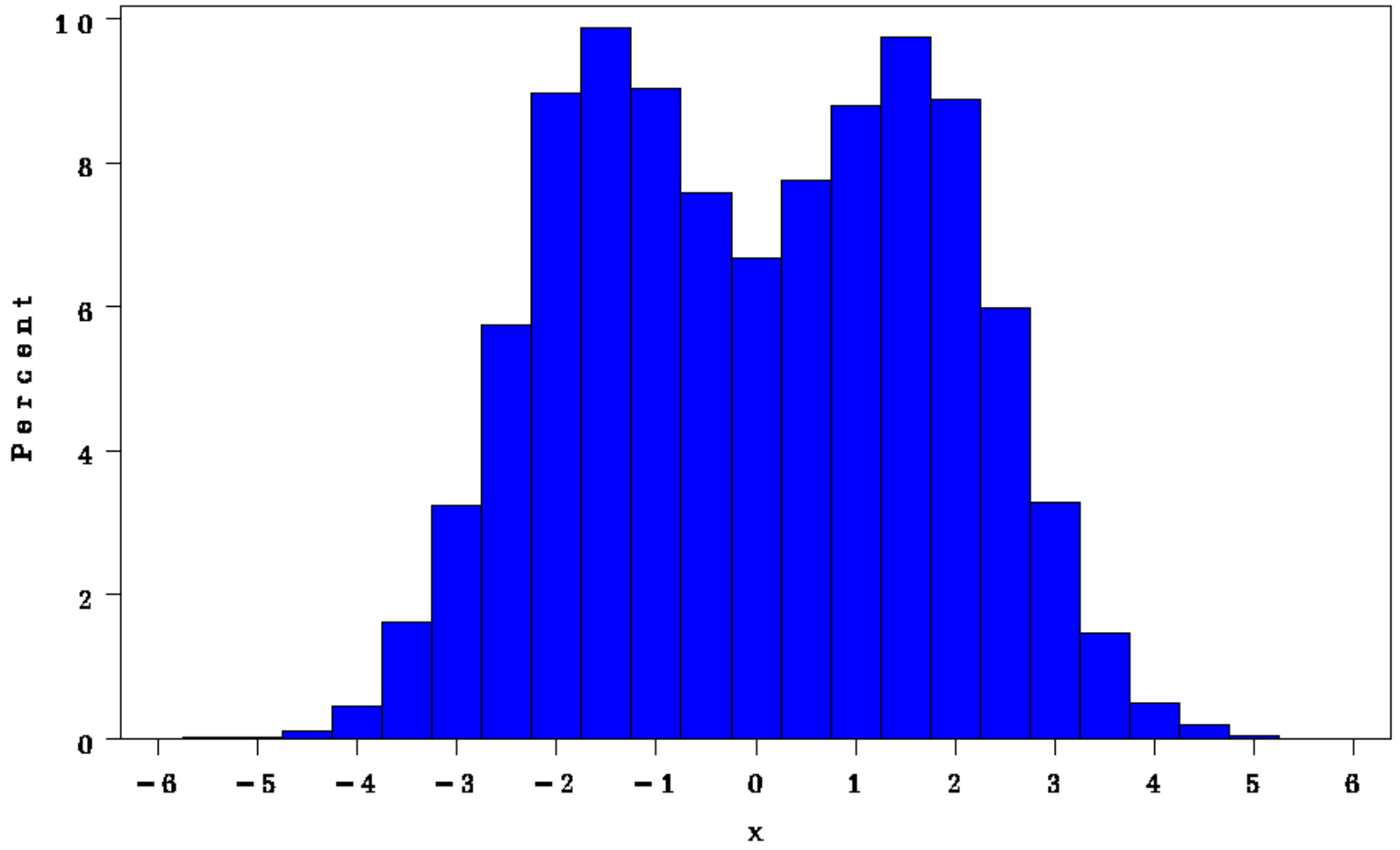
Symmetricity



Symmetric

Unimodal

Modality

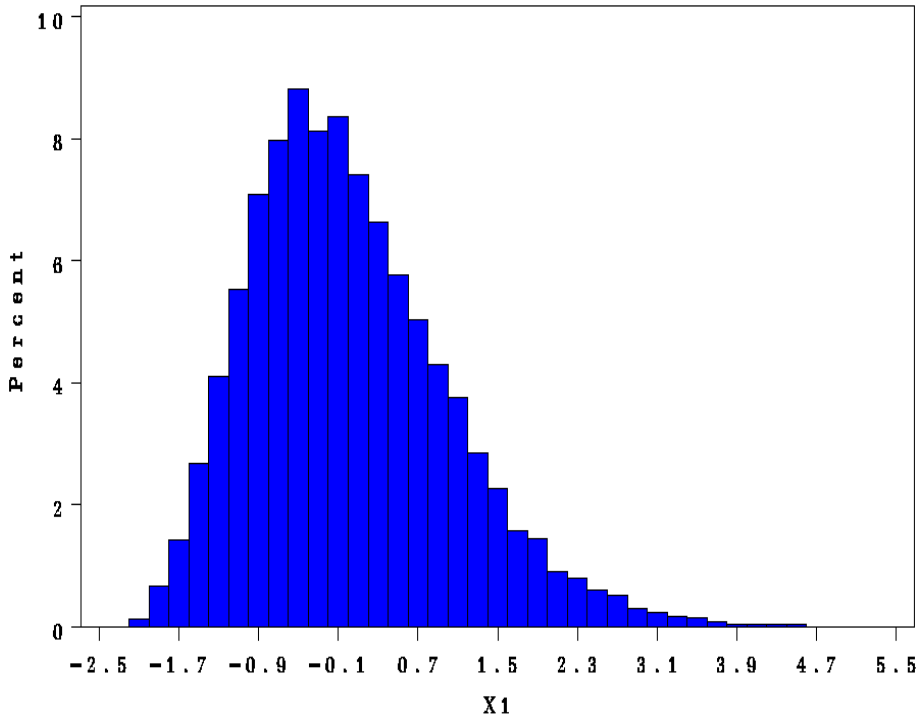


Symmetric

Bimodal

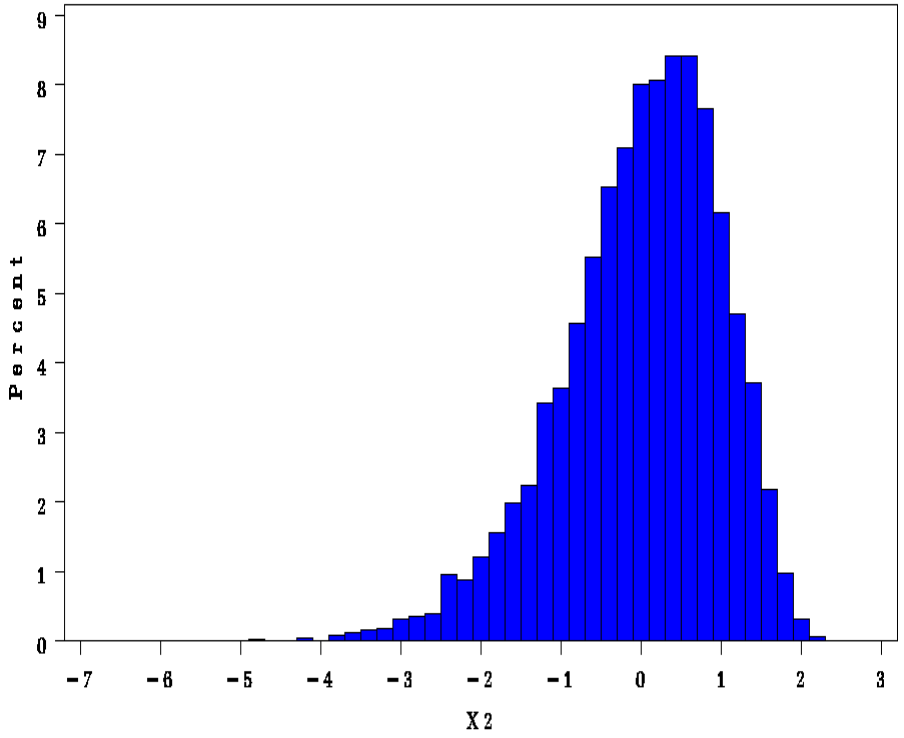
Skewness

(C)



Positive Skew

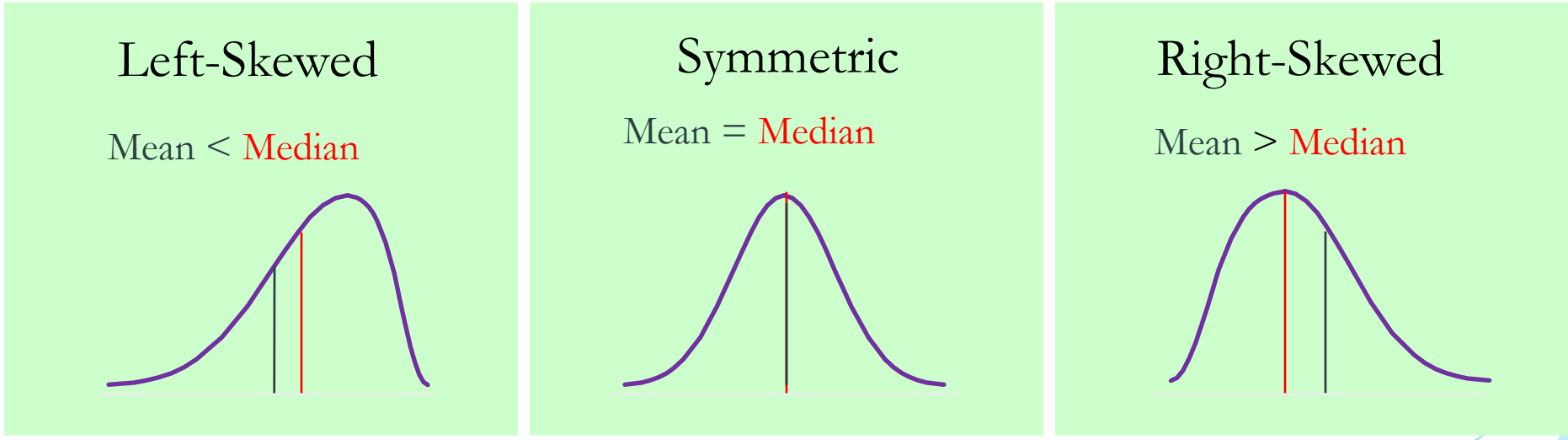
(D)



Negative Skew

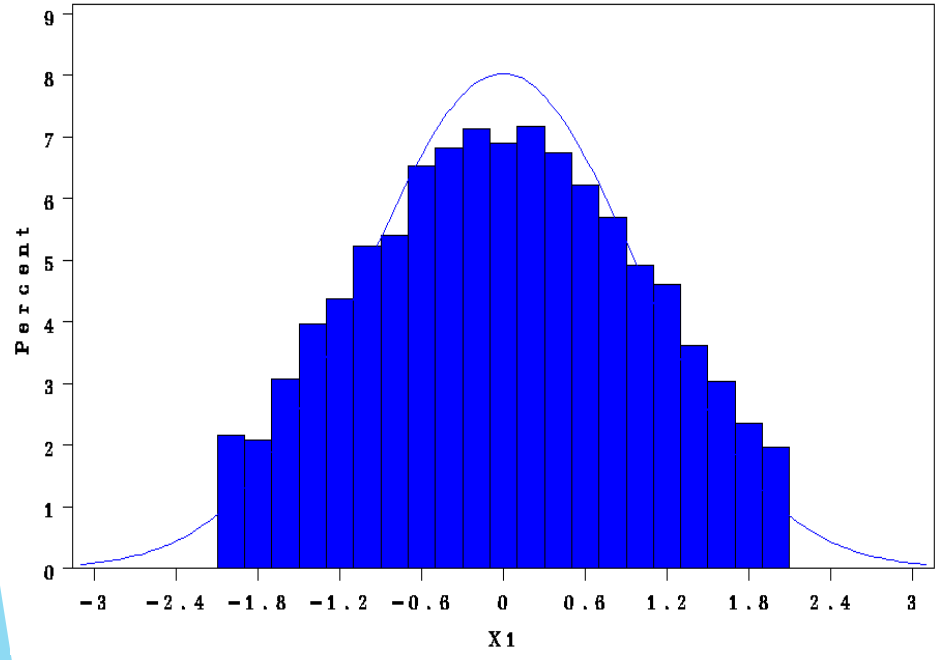
Shape of a distribution

- Describes how data are distributed
- Measures of shape
  - Symmetric or skewed



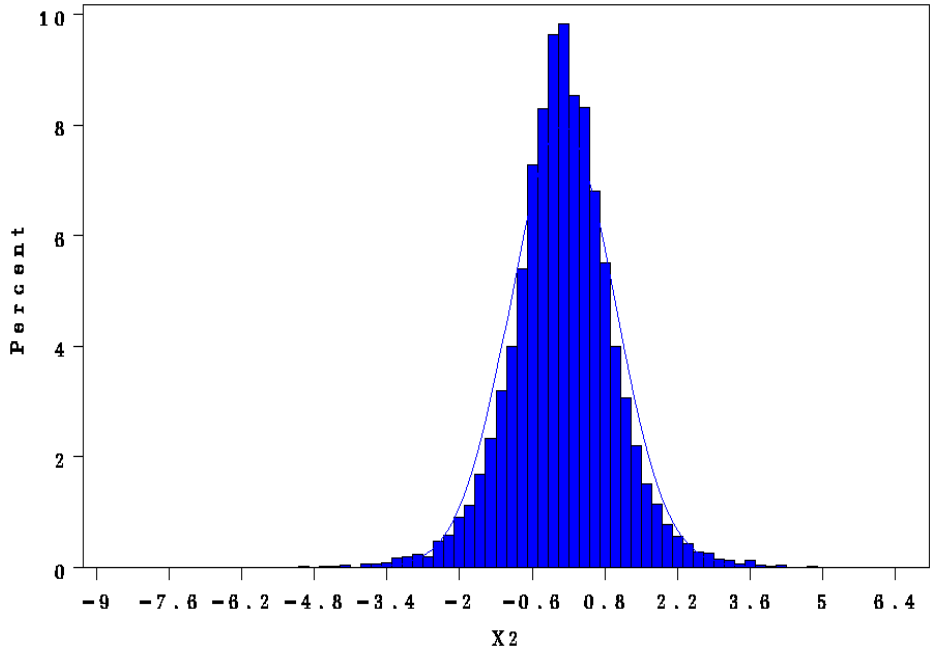
Kurtosis

(E)



Platykurtic

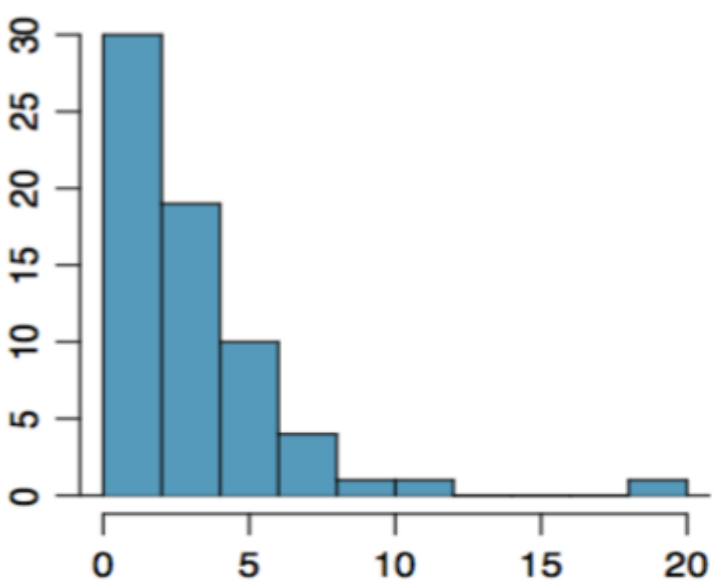
(F)

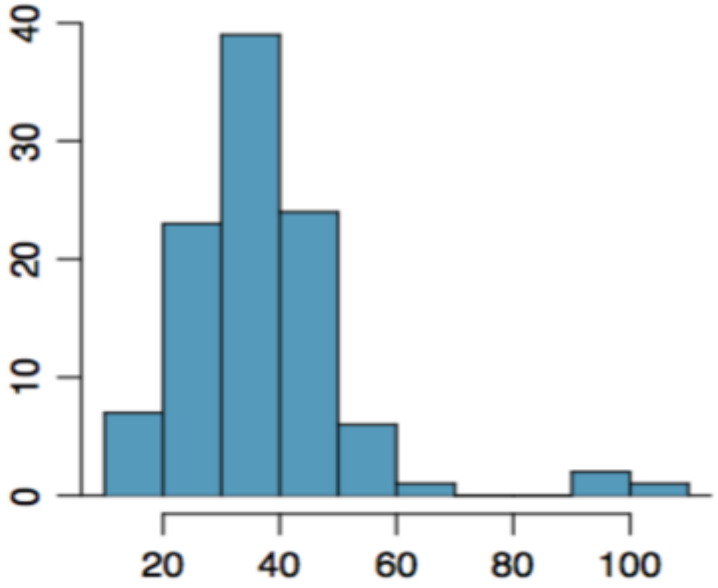


Leptokurtic

Outliers/Unusual Observations

Are there any unusual observations or potential outliers?

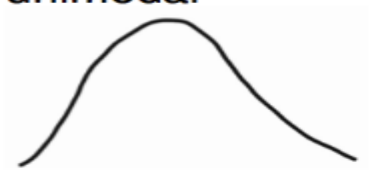




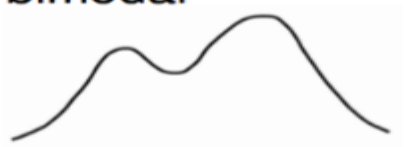
Commonly Observed Shapes

Modality

unimodal

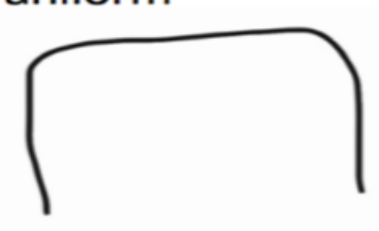


bimodal



multimodal

uniform



Skewness

right skew

left skew

symmetric

Practice

Which of these variables do you expect to be uniformly distributed?

a) Weights of adult females
b) Salaries of a random sample of people from North Carolina
c) House prices
d) Birthdays of classmates (day of the month)

Practice

Which of these variables do you expect to be uniformly distributed?

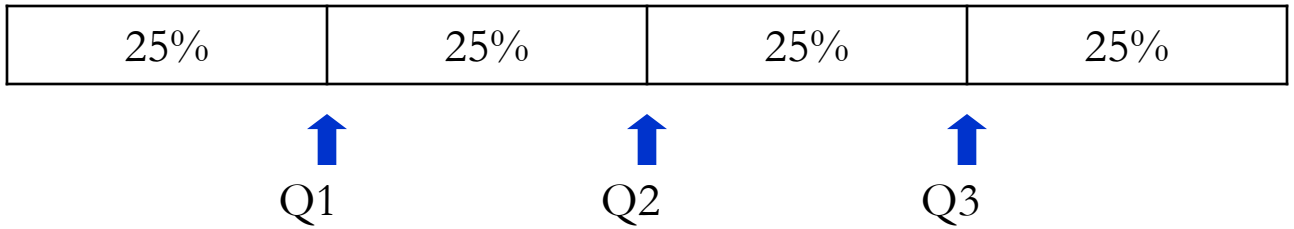
a) Weights of adult females
b) Salaries of a random sample of people from North Carolina
c) House prices
d) *Birthdays of classmates (day of the month)*

Measures of Relative Standing

- Quartiles
- Interquartile range
- Z score

Quartiles

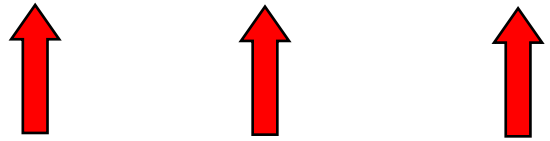
- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- 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% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartiles (Example)

Sample Data in Ordered Array: 7 8 11 12 13 16 16 17 18 21 22



(n = 11) (n+1)/4

Q_1 is in the $(11+1)/4 = 3^{rd}$ position of the ranked data,

so $Q_1 = 11$

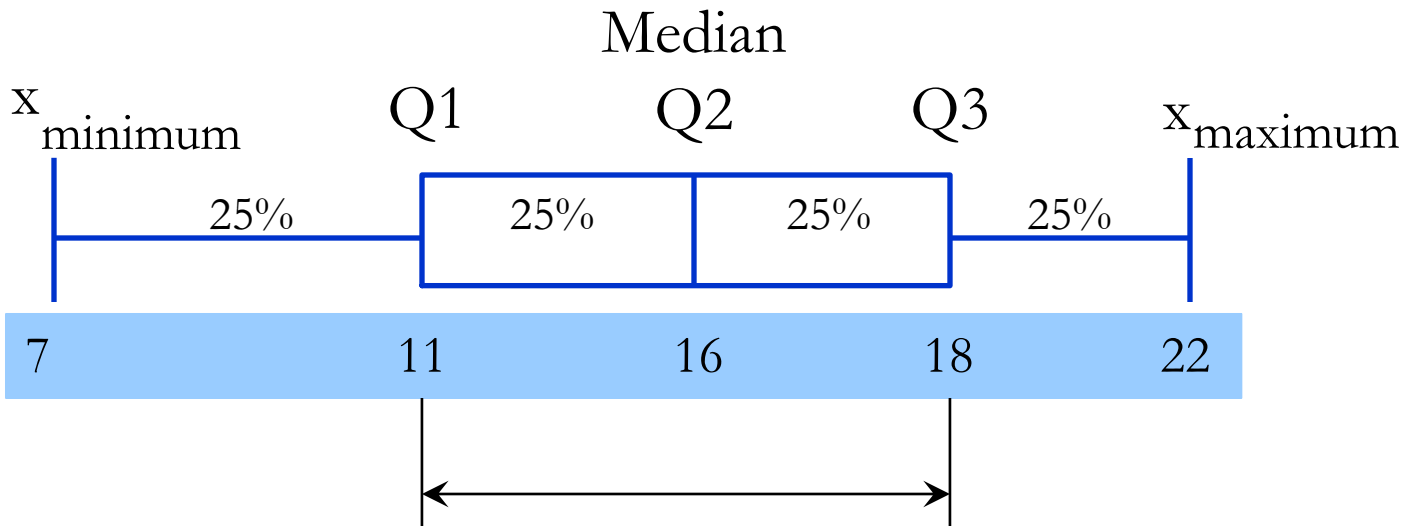
Q_2 is in the $2(11+1)/4 = 6^{th}$ position of the ranked data,

so Q_2 = median = 16

Q_3 is in the $3(11+1)/4 = 9^{th}$ position of the ranked data,

so $Q_3 = 18$

Interquartile range (IQR)



Interquartile range (IQR)

$= Q_3 - Q_1$

$= 18 - 11 = 7$

Quartiles Calculation (based on textbook)

- Order the n observations from smallest to largest
- Calculate the product *np*
- n = sample size
- p = 0.25 for 25th percentile
  = 0.5 for 50th percentile
  = 0.75 for 75th percentile
- If *np* is not an integer, round it up to the next integer and find the corresponding ordered value
- If *np* is an integer (say k), take the average of the kth and k+1th observations

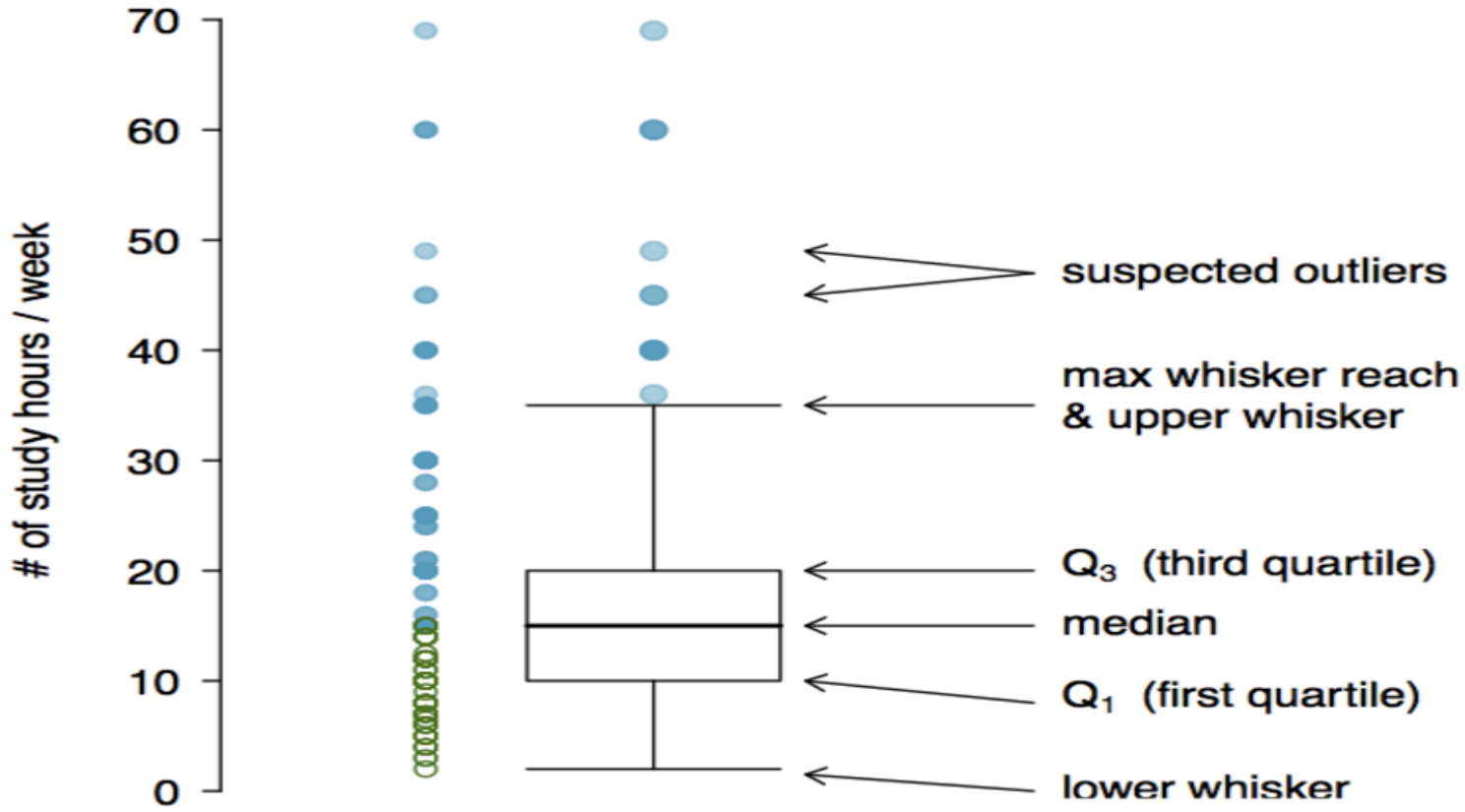
Box Plot

- Graphical procedure that contains information about the mean, median, interquartile range, and unusual and/or extreme observations.

Box Plot

The Box (or Box Plot) represents the middle 50% of the data.

The Whiskers represent the general max and min of the data.



Box Plot

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

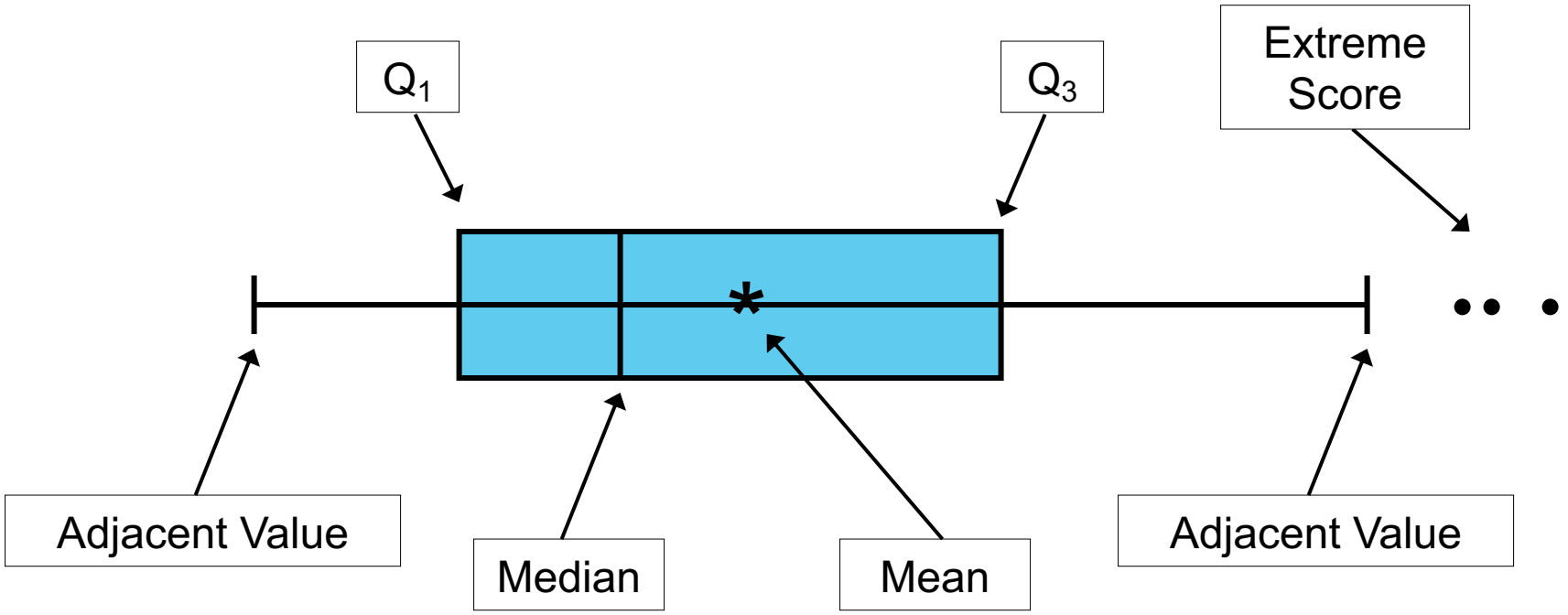
IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

A potential outlier is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Box Plot


Q_1
Q_3
Extreme Score
Adjacent Value
Median
Mean
Adjacent Value

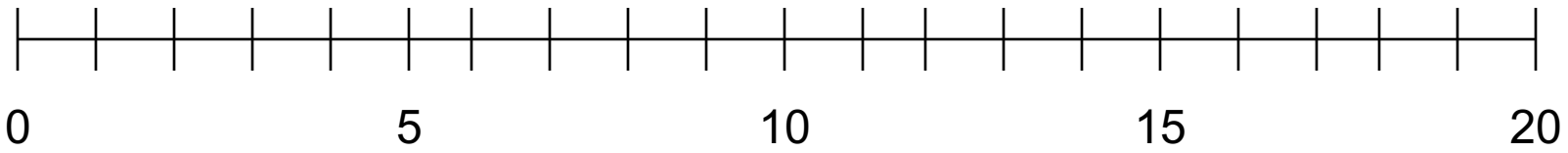
Outliers

- Important to look at outliers
  - Identify extreme skew in the distribution
  - Identify data collection and entry errors
  - Provide greater insights into interesting data features

Example

- Create a Box Plot for the following data

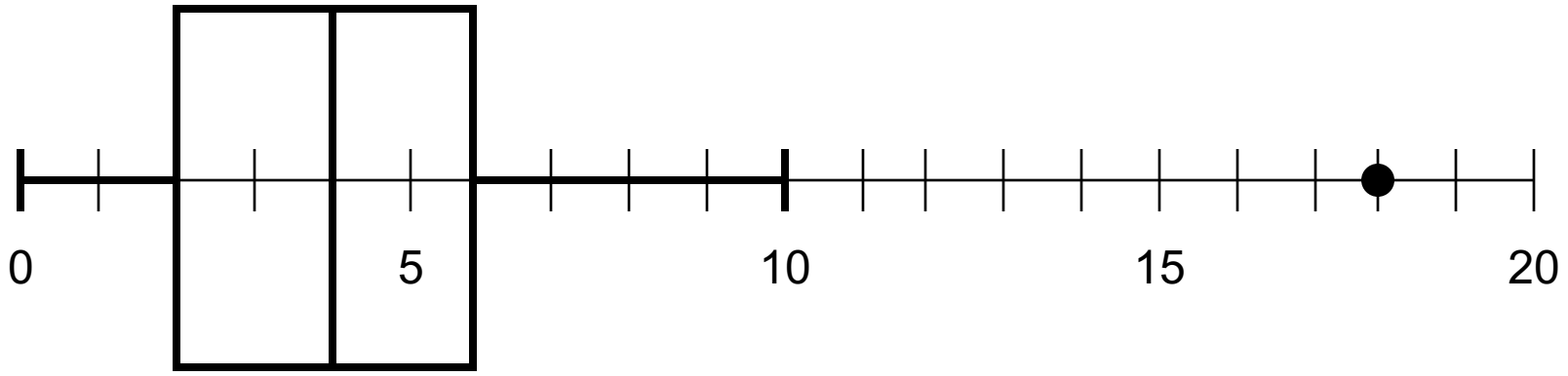
  0, 0, 0, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 6, 9, 9, 10, 18



Example

0, 0, 0, 1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 6, 9, 9, 10, 18

- Median = 4, Q1= 2, Q3 = 6
- Lower Fence = Q1 – 1.5 * (Q3 – Q1) = 2 – 1.5* 4 = -4
  - Adjacent Value = 0
- Upper Fence = Q3 + 1.5 * (Q3 – Q1) = 6 + 1.5 * 4 = 12
  - Adjacent Value = 10



Z-Score

- A measure of distance from the mean (for example, a Z-score of 2.0 means that a value is 2.0 standard deviations from the mean).
- The difference between a value and the mean, divided by the standard deviation.
- A Z-score above 3.0 or below -3.0 is considered as an outlier.

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

Z-Score Example

- If the mean is 15.0 and the standard deviation is 3.0, what is the Z score for the value 18?

$$Z = \frac{X - \bar{X}}{S} = \frac{18 - 15}{3.0} = 1$$

- The value 18 is 1.0 standard deviations above the mean
- (A negative Z-score would mean that a value is less than the mean)

Practice Problems from Textbook

- 2.1-2.10

Descriptive Measures in Excel

Example

Find the mean, median, mode, range and standard deviation for the following list of values:

13, 18, 13, 14, 13, 16, 14, 21, 13

Mean	
Median	
Mode	
Range	
Standard Deviation	

Example

Find the mean, median, mode, range and standard deviation for the following list of values:

13, 18, 13, 14, 13, 16, 14, 21, 13

Mean	15
Median	14
Mode	13
Range	8
Standard Deviation	2.828

Example

Find the mean, median, Q1, Q3, IQR for the following data and represent it on a boxplot

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Example

Find the mean, median, Q1, Q3, IQR for the following data and represent it on a boxplot

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Hand Calculation
Q1 = 278
Q3 = 330
IQR = 52

Excel
Q1 = 279.5
Q3 = 328.25
IQR = 48.75

Mean = 305.58
Median = 304.5

Numerical Methods for Describing Qualitative Data

- **Frequency Distribution Table**
- **Contingency Table**

Frequency Distribution Table

- **Category frequency:** number of observations that fall in a given category.
- **Category relative frequency:** the proportion of the number of observations that fall in a given category.

Frequency Distribution Table - Example

TABLE 2.1 Summary Frequency Table for Cause of Energy-Related Fatal Accidents

Category *(Cause)*	Frequency *(Number of Accidents)*	Relative Frequency *(Proportion)*
Coal mine collapse	7	.156
Dam failure	4	.089
Gas explosion	28	.622
Lightning	1	.022
Nuclear reactor	1	.022
Oil fire	4	.089
Totals	45	1.000

Source: "Safety of Nuclear Power Reactors." *Nuclear Issues Briefing Paper 14*, November 2004.

Contingency Table

- Summarize two or more qualitative/categorical variables at the same time
- Looking at gender and promotions data at a workplace

		Decision		
		promoted	not promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

Contingency Table

		Decision		
		promoted	not promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

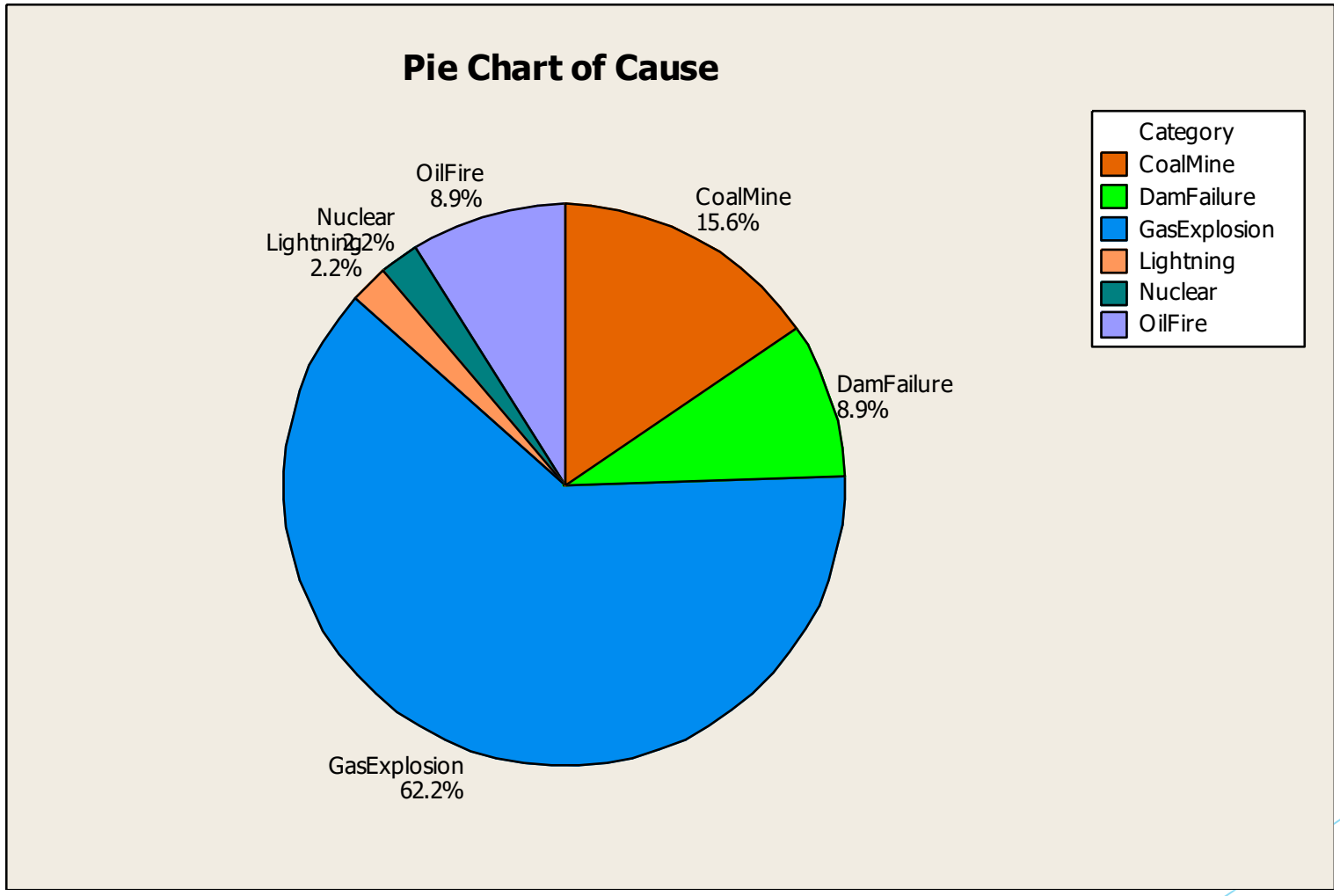
Gender	Total
Male	24
Female	24

Promotion	Total
Promoted	35
Not Promoted	13

Graphical Methods for Describing Qualitative Data

- Pie Chart
- Bar Chart
- Pareto Diagram

Pie Chart


Pie Chart of Cause
Category
CoalMine
DamFailure
GasExplosion
Lightning
Nuclear
OilFire
CoalMine
15.6%
DamFailure
8.9%
GasExplosion
62.2%
Lightning
2.2%
Nuclear
2.2%
OilFire
8.9%

Bar Chart

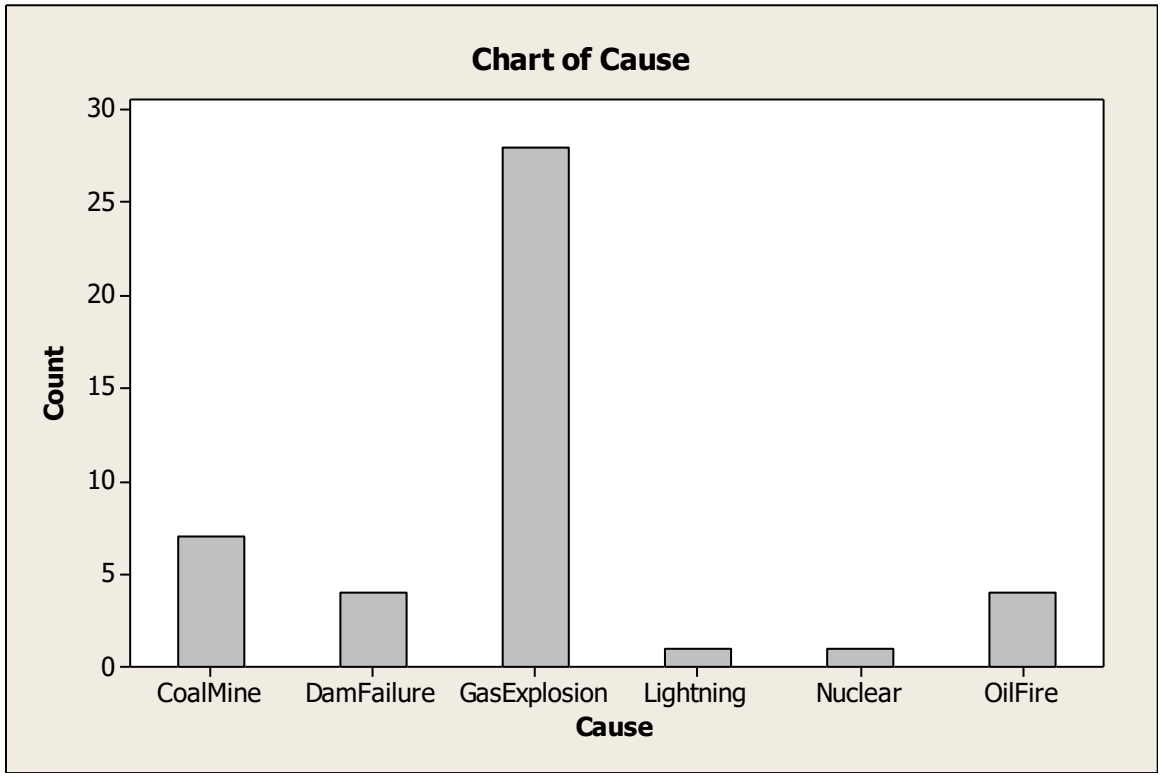
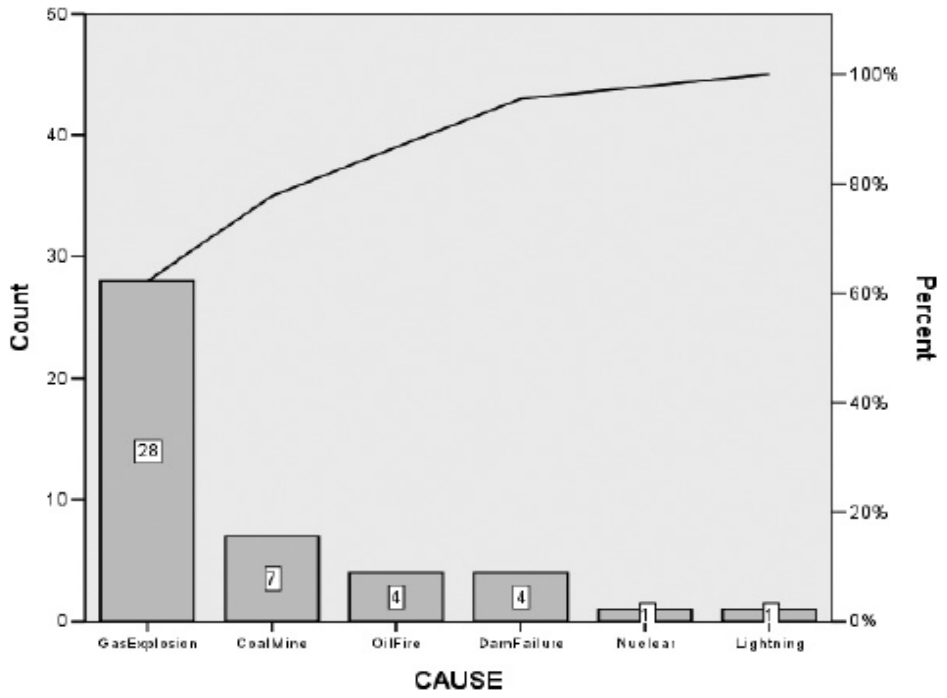


Chart of Cause
Count
30
25
20
15
10
5
0
CoalMine
DamFailure
GasExplosion
Lightning
Nuclear
OilFire
Cause

Pareto Diagram



Cumulative