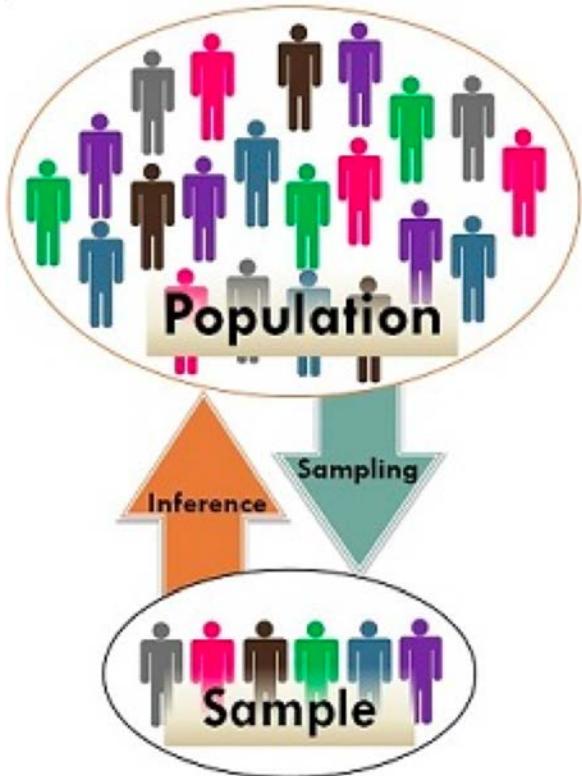


## Descriptive Statistical Measures

# Learning objectives

- Explain the difference between Populations and Samples
- Understand and be able to distinguish and apply the different measures:
  - Measures of Location (Mean, Median, Mode),
  - Measures of Dispersion (Range, Variance, Standard Deviation, Chebyshev's Theorem, Coefficient of Variation)
  - Measures of Shape (Skewness, Kurtosis)
  - Measures of Association (Covariance and Correlation)
- Be able to identify outliers

# Populations and Samples



- ❖ Population - all items of interest for a particular decision or investigation
  - *all* married drivers over 25 years old
  - *all* subscribers to Netflix
- ❖ Sample - a subset of the population
  - a list of married drivers over 25 years old who bought a new car in the past year
  - a list of individuals who rented a comedy from Netflix in the past year
- ❖ Purpose of sampling is to obtain sufficient information to draw a valid inference about a population

# Measures of Location – Mean

For a population of size  $N$ :

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

For a sample of  $n$  observations:

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

*Mean* is also commonly known as *average*

# Measures of Location – Mean

Example: Computing **Mean Cost** per Order (Using *Purchase Orders* data)

- Using formula:

$$\text{Mean} = \$2,471,760/94 \\ = \underline{\$26,295.32}$$

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

A	B	C	D	E	F	G	H	I	J	
1	Purchase Orders									
2	Supplier	Order No	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Month)	Order Date	Arrival Date
4	Spacetime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25	10/10/11	10/18/11
5	Steelpin Inc.	A0115	5319	Shielded Cable/ft.	\$ 1.10	17,500	\$ 19,250.00	30	08/20/11	08/31/11
6	Steelpin Inc.	A0123	4312	Bolt-nut package	\$ 3.75	4,250	\$ 15,937.50	30	08/25/11	09/01/11
7	Steelpin Inc.	A0204	5319	Shielded Cable/ft.	\$ 1.10	16,500	\$ 18,150.00	30	09/15/11	10/05/11
8	Steelpin Inc.	A0205	5677	Side Panel	\$195.00	120	\$ 23,400.00	30	11/02/11	11/13/11
9	Steelpin Inc.	A0207	4312	Bolt-nut package	\$ 3.75	4,200	\$ 15,750.00	30	09/01/11	09/10/11
10	Alum Sheeting	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11
11	Alum Sheeting	A0433	5417	Control Panel	\$255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
12	Alum Sheeting	A0443	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00	30	08/08/11	08/14/11
13	Alum Sheeting	A0446	5417	Control Panel	\$255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
14	Spacetime Technologies	A0533	9752	Gasket	\$ 4.05	1,500	\$ 6,075.00	25	09/20/11	09/25/11
15	Spacetime Technologies	A0555	6489	O-Ring	\$ 3.00	1,100	\$ 3,300.00	25	10/05/11	10/10/11

# Measures of Location – Median

~ middle value of the data when arranged from least to greatest

Example: Finding the Median Cost per Order (*Purchase Orders* data)

Sort the data in column B.

Since  $n = 94$ ,

Median = \$15,656.25

(average of 47<sup>th</sup> &  
48<sup>th</sup> observations)

	A	B	C	D
1	Rank	Cost per order		
2		\$68.75		
3		\$82.50		
4		\$375.00		
5		\$467.50		
6		\$525.00		
44	43	\$13,650.00		
45	44	\$14,910.00		
46	45	\$14,910.00		
47	46	\$15,087.50		
48	47	\$15,562.50		\$15,562.50
49	48	\$15,750.00		\$15,750.00
50	49	\$15,937.50	Average	\$15,656.25
51	50	\$16,276.75		

NOTE: If  $n$  is odd number, then median is the value of the middle observation

# Using R to compute mean and median

```
mean(x, trim = 0, na.rm = FALSE, ...)
```

## Arguments

- x** An R object. Currently there are methods for numeric/logical vectors and [date](#), [date-time](#) and [time interval](#) objects. Complex vectors are allowed for `trim = 0`, only.
- trim** the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of `trim` outside that range are taken as the nearest endpoint.
- na.rm** a logical value indicating whether NA values should be stripped before the computation proceeds.

```
median(x, na.rm = FALSE, ...)
```

## Arguments

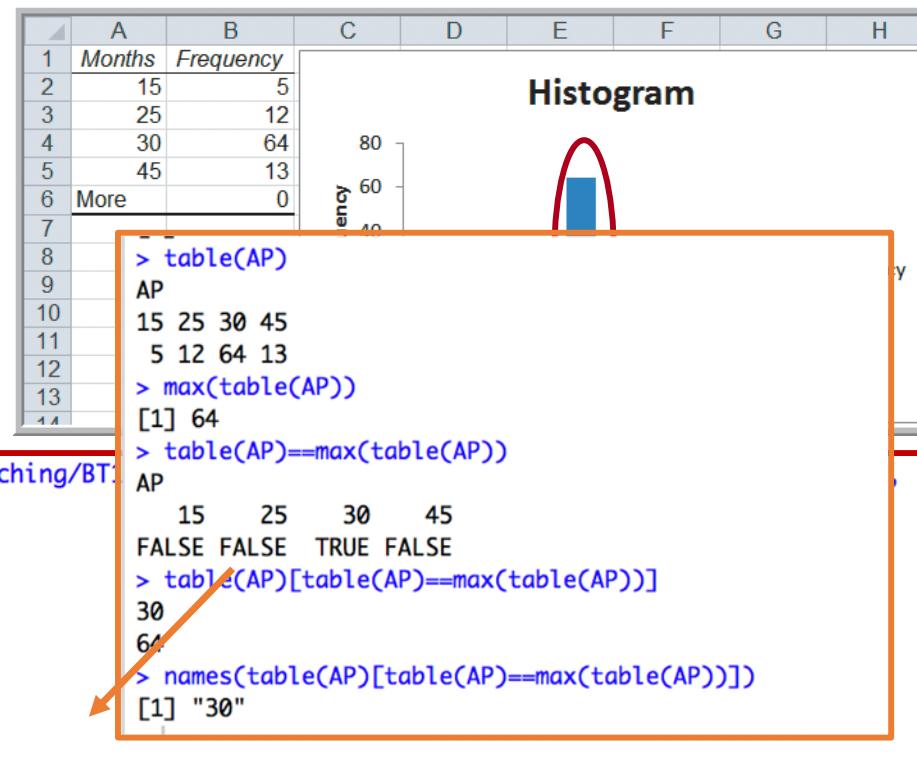
- x** an object for which a method has been defined, or a numeric vector containing the values whose median is to be computed.
- na.rm** a logical value indicating whether NA values should be stripped before the computation proceeds.

# Measures of Location – Mode

~ observation that occurs most often or, for grouped data, the group with the greatest frequency.

Eg for observation data: Finding the Mode of A/P terms  
(*Purchase Orders* data)

- Mode of A/P terms:  
= 30 months



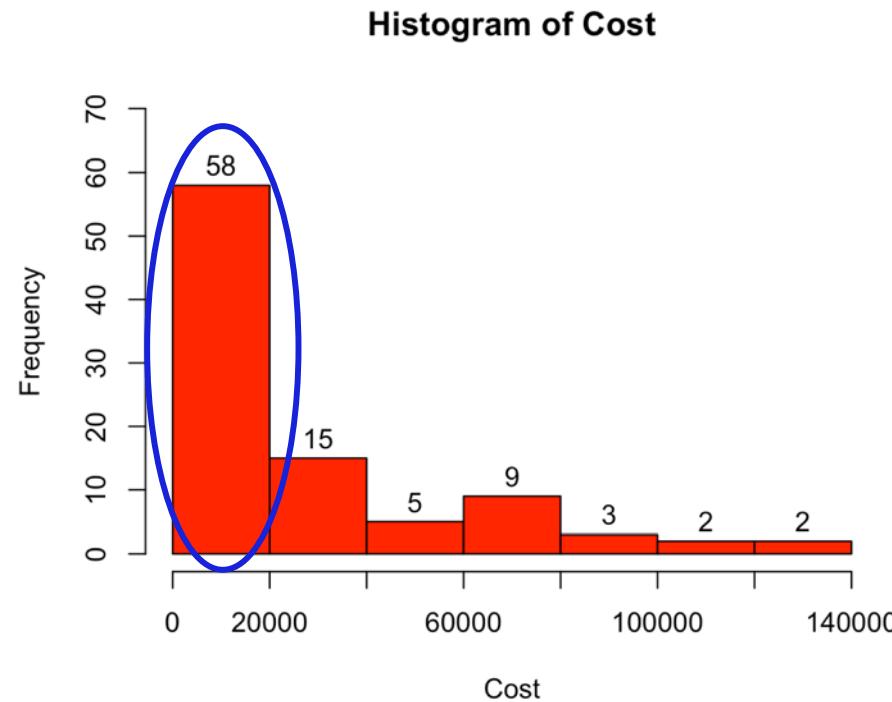
```
> Purchase_Orders <- read_excel("~/Dropbox/Teaching/BT/..."))
+   sheet = "Data", col_types = c("text",
+     "text", "text", "text", "numeric",
+     "numeric", "numeric", "numeric",
+     "date", "date"), skip = 2)
> View(Purchase_Orders)
> AP<-Purchase_Orders$`A/P Terms (Months)`
> names(table(AP))[table(AP)==max(table(AP))]
[1] "30"
```

# Measures of Location – Mode

~ observation that occurs most often or, for grouped data, the group with the greatest frequency.

Eg for grouped data: Finding the Mode of Cost per order (*Purchase Orders* data)

- Mode is the group between \$0 and \$20,000



# Measures of Location: Application

## Problem: Quoting Computer Repair Times

Data set (Computer Repair Times) includes 250 repair times for customers.

- What repair time would be reasonable to quote to a new customer?

```
> df8<-Computer_Repair_Times  
# Use R functions for Mean & Median  
  
> mean(df8$`Repair Time (Days)`)  
[1] 14.912  
> median(df8$`Repair Time (Days)`)  
[1] 14
```

# Compute Mode

```
> x<-df8$`Repair Time (Days)`  
> names(table(x))[table(x)==max(table(x))]  
[1] "12" "15"
```

# Use Table function to obtain frequencies for each value of X

```
> table(x)
```

x

5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	25	26	27	29	31	34	36	37	38	39	40
1	2	5	12	14	19	19	23	22	20	23	18	16	9	9	8	5	4	4	3	2	2	2	1	1	2	1	1	1	1

	Sample	Repair Time (Days)
1	1	18
2	2	15
3	3	17
4	4	9
5	5	37
6	6	15

	Sample	Repair Time (Days)
241	241	21
242	242	7
243	243	12
244	244	16
245	245	23
246	246	18
247	247	31
248	248	6
249	249	17
250	250	13

Mean repair time is about 15 days

Median repair time: 2 weeks

Mode is 12 and 15 days

# Measures of Location: Application

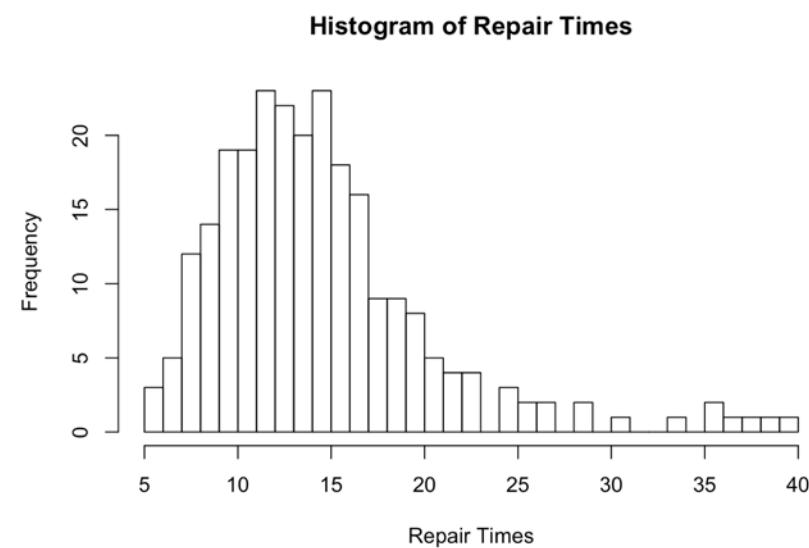
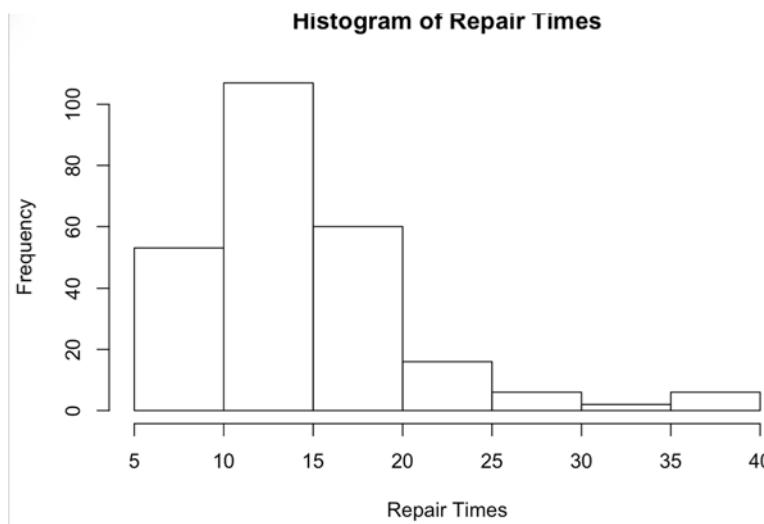
## Problem: Quoting Computer Repair Times

Data set (Computer Repair Times) includes 250 repair times for customers.

- **What repair time would be reasonable to quote to a new customer?**
- Mean repair time is about 15 days
- Median repair time: 2 weeks
- Mode is 12 and 15 days

```
hist(x, main="Histogram of Repair Times", xlab="Repair Times")
```

```
hist(x, breaks=36, xlim=range(5,40), main="Histogram of Repair Times", xlab="Repair Times")
```



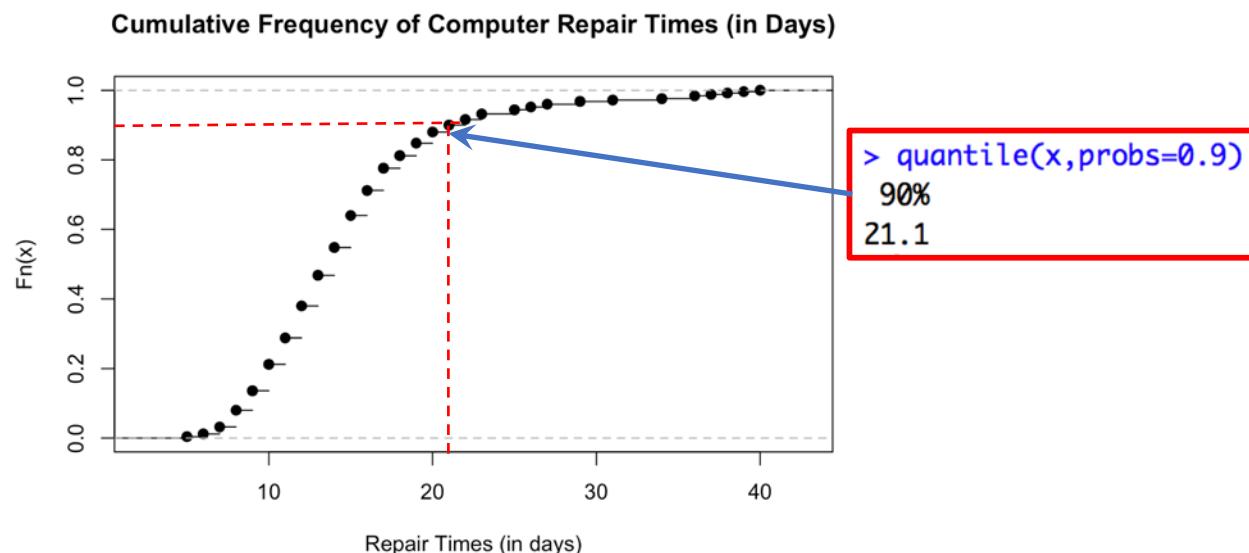
# Measures of Location: Application

Problem: Quoting Computer Repair Times

Data set (Computer Repair Times) includes 250 repair times for customers.

- What repair time would be reasonable to quote to a new customer?
- Mean repair time is about 15 days
- Median repair time: 2 weeks
- Mode is 12 and 15 days

```
plot(ecdf(x),main ="Cumulative Frequency of Computer Repair Times (in Days)", xlab="Repair Times (in days)")
```



# Measures of Dispersion

- Dispersion refers to the degree of variation (numerical spread or compactness) in the data
  - Range is the difference between the maximum and minimum data values
  - Interquartile range difference between the first and third quartiles ( $Q_3 - Q_1$ ) (uses middle 50% data)
  - Variance is an average of squared deviations from mean (uses all data values) `var()`
  - Standard Deviation is the square root of the variance `sd()`

x is the computer  
repair times variable

```
> summary(x)
   Min. 1st Qu. Median     Mean 3rd Qu.    Max. 
   5.00   11.00  14.00  14.91  17.00  40.00 
                                         Range=?  
                                         IQR =?
```

install.packages (psych)  
library (psych)

```
> describe(x)
      vars   n  mean    sd median trimmed  mad min max range skew kurtosis    se
X1      1 250 14.91  5.96      14  14.12  4.45     5  40    35 1.67     3.92  0.38
```

# Measures of Dispersion – Variance

~ average of squared deviations from mean

## Computing the Variance

- For a population:

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

- For a sample:

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

- Note the difference in denominators

If sample data is also population data, then n=N, to compute population variance:

$$[(N-1)/N]*\text{var}(X)$$

In R, sample variance is `var(X)`

# Measures of Dispersion – Standard Deviation

~ square root of the variance  
(popular measure of risk)

## Computing the Standard Deviation

- For a population:

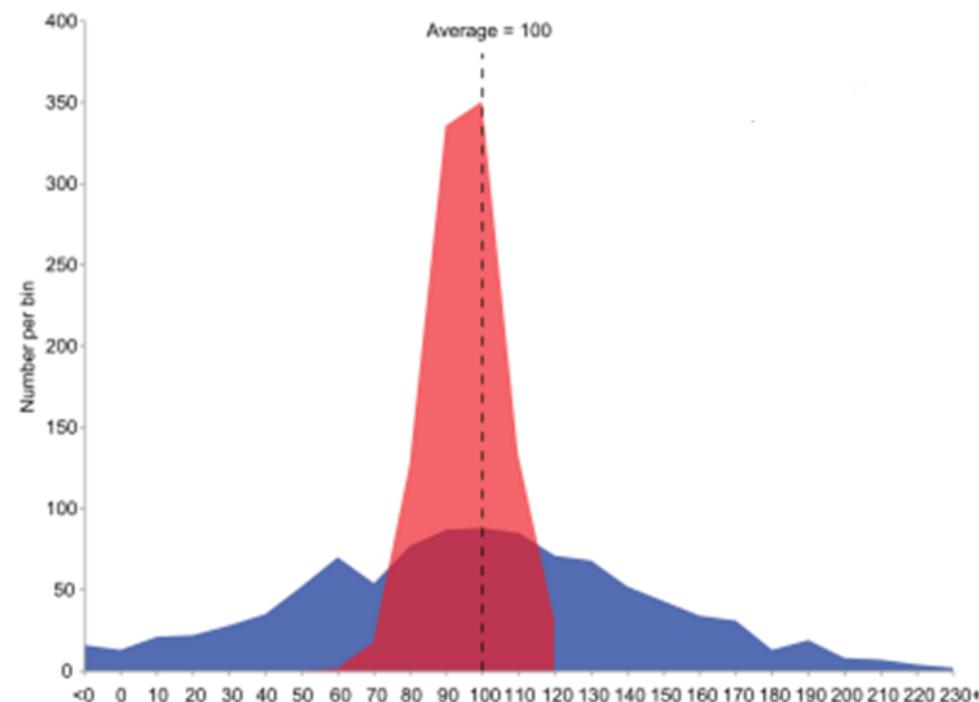
$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{Sqrt(((N-1)/N)*var(X))}$$

- For a sample:

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

# Measures of Dispersion – Standard Deviation

- Which has a **higher** standard deviation?



Source: Wikipedia (Standard Deviation)

# Measures of Dispersion - Application

Mean & Standard Deviation of Closing Stock Prices

Intel (INTC):

Mean = \$18.81

Stdev. = \$0.50

General Electric (GE):

Mean = \$16.19

Stdev. = \$0.35

Whose risk may be higher?

	A	B	C	D	E	F
1	Closing Stock Prices					
2						
3	Date	IBM	INTC	CSCO	GE	DJ Industrials Index
4	9/3/2010	\$127.58	\$18.43	\$21.04	\$15.39	10447.93
5	9/7/2010	\$125.95	\$18.12	\$20.58	\$15.44	10340.69
6	9/8/2010	\$126.08	\$17.90	\$20.64	\$15.70	10387.01
7	9/9/2010	\$126.36	\$18.00	\$20.61	\$15.91	10415.24
8	9/10/2010	\$127.99	\$17.97	\$20.62	\$15.98	10462.77
9	9/13/2010	\$129.61	\$18.56	\$21.26	\$16.25	10544.13
10	9/14/2010	\$128.85	\$18.74	\$21.45	\$16.16	10526.49
11	9/15/2010	\$129.43	\$18.72	\$21.59	\$16.34	10572.73
12	9/16/2010	\$129.67	\$18.97	\$21.93	\$16.23	10594.83
13	9/17/2010	\$130.19	\$18.81	\$21.86	\$16.29	10607.85
14	9/20/2010	\$131.79	\$18.93	\$21.75	\$16.55	10753.62
15	9/21/2010	\$131.98	\$19.14	\$21.64	\$16.52	10761.03
16	9/22/2010	\$132.57	\$19.01	\$21.67	\$16.50	10739.31
17	9/23/2010	\$131.67	\$18.98	\$21.53	\$16.14	10662.42
18	9/24/2010	\$134.11	\$19.42	\$22.09	\$16.66	10860.26
19	9/27/2010	\$134.65	\$19.24	\$22.11	\$16.43	10812.04
20	9/28/2010	\$134.89	\$19.51	\$21.86	\$16.44	10858.14
21	9/29/2010	\$135.48	\$19.24	\$21.87	\$16.36	10835.28
22	9/30/2010	\$134.14	\$19.20	\$21.90	\$16.25	10788.05
23	10/1/2010	\$135.64	\$19.32	\$21.91	\$16.36	10829.68

# Measures of Dispersion

## Chebyshev's Theorem

- For any data set, the proportion of values that lie within  $k$  ( $k > 1$ ) standard deviations of the mean is *at least*  $1 - 1/k^2$

Substituting values of  $k$ , we get:

- For  $k = 2$ : at least  $\frac{3}{4}$  or **75%** of the data lie within **two standard deviations** of the mean
- For  $k = 3$ : at least  $\frac{8}{9}$  or **89%** of the data lie within **three standard deviations** of the mean

Why is this useful?

- Able to use mean and standard deviation to find percentage of total observations that fall within a given interval about the mean

## Example: For Cost per order data in Purchase orders database

Purchase Orders									
Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
Alum Sheeting	Aug11002	1243	Airframe fasteners	\$ 4.25	10,000	\$ 42,500.00	30	08/08/11	08/14/11
Alum Sheeting	Sep11002	5417	Control Panel	\$ 255.00	406	\$ 103,530.00	30	09/01/11	09/10/11
Alum Sheeting	Sep11008	1243	Airframe fasteners	\$ 4.25	9,000	\$ 38,250.00	30	09/05/11	09/12/11
Alum Sheeting	Oct11016	1243	Airframe fasteners	\$ 4.25	10,500	\$ 44,625.00	30	10/10/11	10/17/11
Alum Sheeting	Oct11022	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11
Alum Sheeting	Oct11026	5417	Control Panel	\$ 255.00	500	\$ 127,500.00	30	10/20/11	10/27/11
Alum Sheeting	Oct11028	5634	Side Panel	\$ 185.00	150	\$ 27,750.00	30	10/25/11	11/03/11
Alum Sheeting	Oct11036	5634	Side Panel	\$ 185.00	140	\$ 25,900.00	30	10/29/11	11/04/11
Durable Products	Aug11008	7258	Pressure Gauge	\$ 90.00	100	\$ 9,000.00	45	08/25/11	08/28/11
Durable Products	Sep11009	7258	Pressure Gauge	\$ 90.00	120	\$ 10,800.00	45	09/05/11	09/09/11
Durable Products	Sep11027	1369	Airframe fasteners	\$ 4.20	15,000	\$ 63,000.00	45	09/25/11	09/30/11
Durable Products	Sep11031	1369	Airframe fasteners	\$ 4.20	14,000	\$ 58,800.00	45	09/27/11	10/03/11
Durable Products	Sep11034	1369	Airframe fasteners	\$ 4.20	10,000	\$ 42,000.00	45	09/29/11	10/04/11
Durable Products	Oct11002	9399	Gasket	\$ 3.65	1,250	\$ 4,562.50	45	10/01/11	10/06/11
Durable Products	Oct11007	9399	Gasket	\$ 3.65	1,450	\$ 5,292.50	45	10/03/11	10/08/11
Durable Products	Oct11009	9399	Gasket	\$ 3.65	1,985	\$ 7,245.25	45	10/05/11	10/11/11

## Example: For Cost per order data in Purchase orders database

- Applying two std dev interval (i.e. k=2)

```
> mean(Cost)
[1] 26295.32
> sd(Cost)
[1] 29842.83
> mean(Cost)-2*(sd(Cost))
[1] -33390.34
> mean(Cost)+2*(sd(Cost))
[1] 85980.98
> (length(Cost[Cost>-33390.34 & Cost<85980.98]))/94
[1] 0.9468085
```

94.68% falls within 2 sd of the mean.

➤ k= 2; 75% of according to Chebyshev's theorem

- Applying three std dev interval (i.e. k=3)

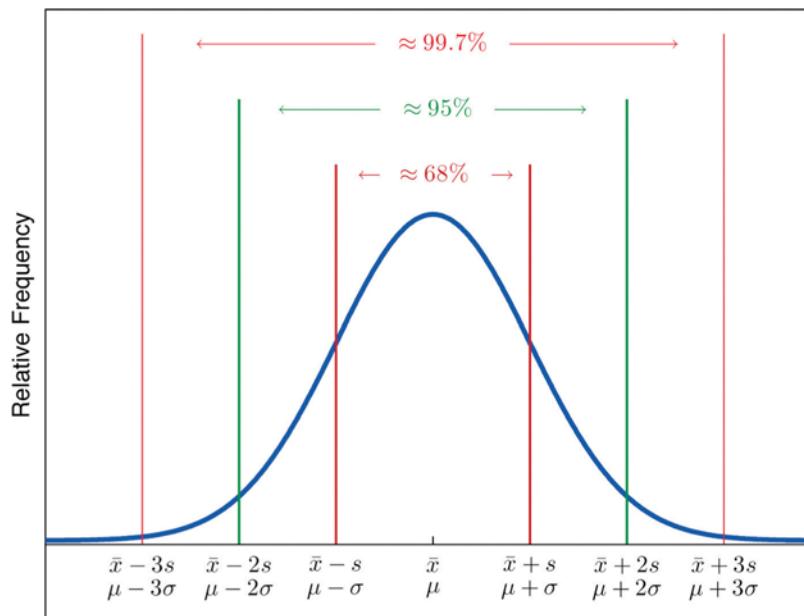
```
> mean(Cost)
[1] 26295.32
> sd(Cost)
[1] 29842.83
> mean(Cost)-3*(sd(Cost))
[1] -63233.17
> mean(Cost)+3*(sd(Cost))
[1] 115823.8
> (length(Cost[Cost>-63233.17 & Cost<115823.8]))/94
[1] 0.9787234
```

97.87% falls within 3 sd of the mean

➤ k=3; 89% according to Chebyshev's theorem

# Measures of Dispersion

- Empirical Rules- For normally distributed data set, the proportion of values that lie within  $k$  ( $k > 1$ ) standard deviations of the mean follows the empirical rules:
  - For  $k = 1$ : about 68% lie within one standard deviations of the mean
  - For  $k = 2$ : about 95% lie within two standard deviations of the mean
  - For  $k = 3$ : about 99.7% lie within three standard deviations of the mean



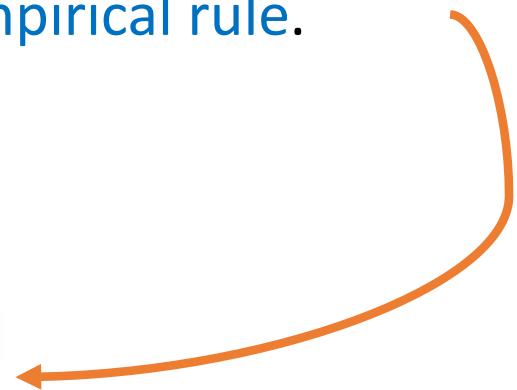
Source: Statistics Libretexts

## Application of Empirical Rule - Process Capability Index

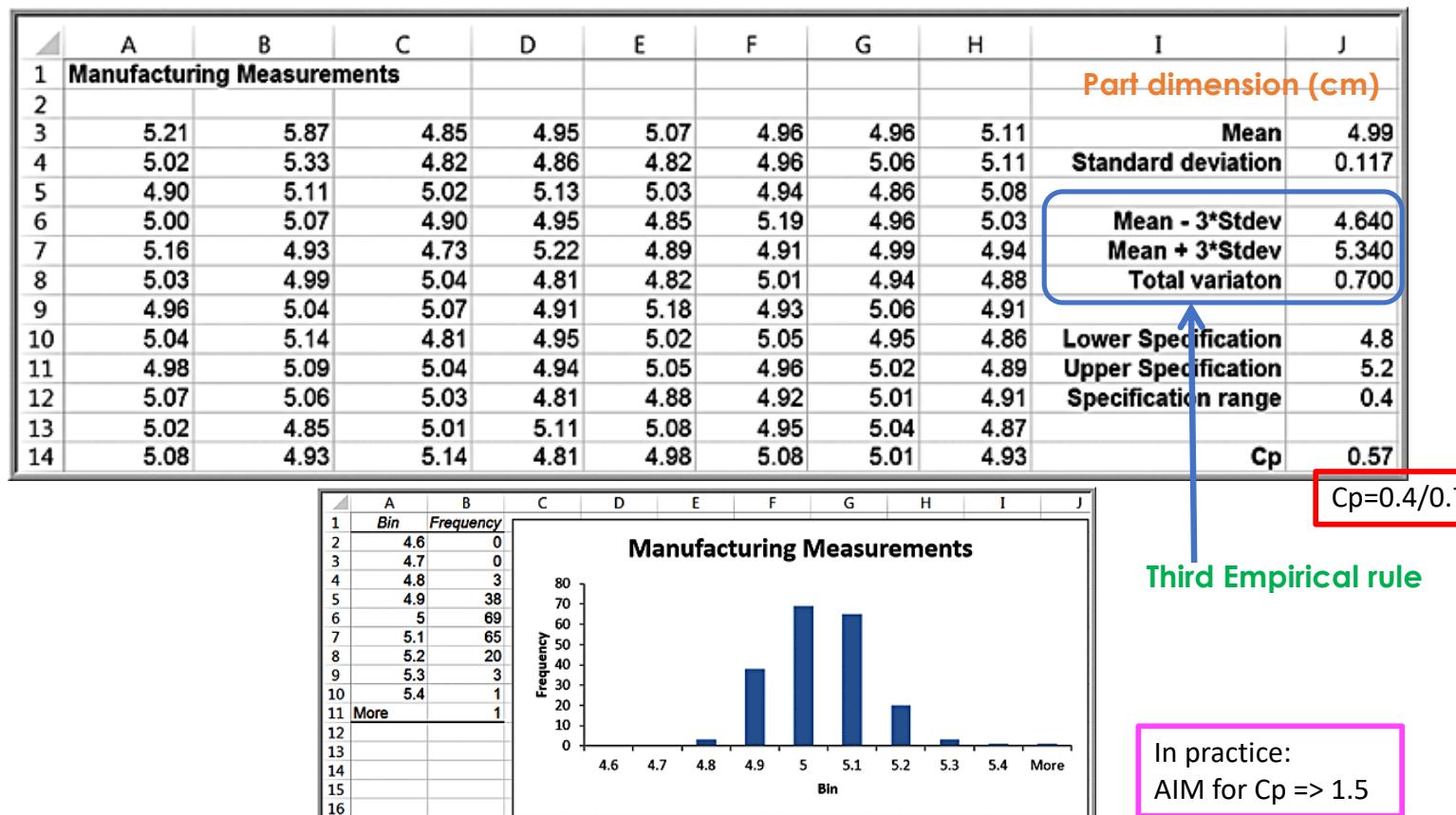
- Process capability index ( $C_p$ ) is a measure of how well a manufacturing process can achieve specifications
- Using a sample of output, measure the dimension of interest, and compute the total variation using the third empirical rule.
- Compare results to specifications using:

$$C_p = \frac{\text{upper specification} - \text{lower specification}}{\text{total variation}}$$

---



## Eg: Using Empirical Rules to Measure the Capability of a Manufacturing Process



$$C_p = 0.4/0.7$$

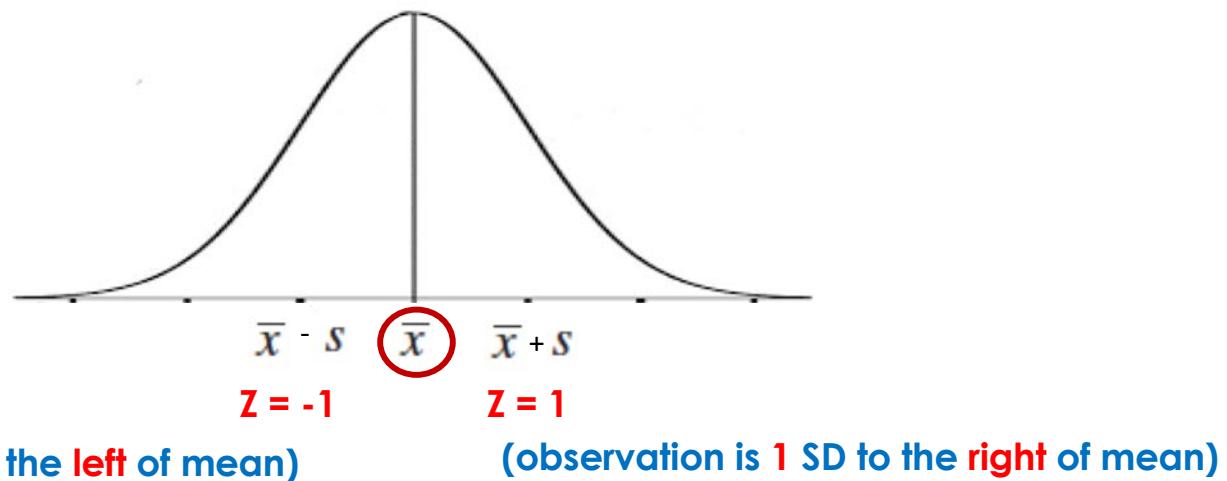
Third Empirical rule

In practice:  
AIM for  $C_p \Rightarrow 1.5$

# Standardized Values

- A **standardized value**, commonly called a **z-score**, provides a **relative measure of the distance an observation is from the mean** (independent of units of measurement)
- z-score for  $i^{\text{th}}$  observation in a data set is:

$$z_i = \frac{x_i - \bar{x}}{s}$$



## EG: Computing z-Scores

- *Purchase Orders Cost per order data*

	A	B	C
1	Observation	Cost per order	z-score
2	x1	\$2,700.00	-0.79
3	x2	\$19,250.00	-0.24
4	x3	\$15,937.50	-0.35
5	x4	\$18,150.00	-0.27
6	x5	\$23,400.00	-0.10
91	x90	\$6,750.00	-0.65
92	x91	\$16,625.00	-0.32
93	x92	\$74,375.00	1.61
94	x93	\$72,250.00	1.54
95	x94	\$6,562.50	-0.66
96			
97	Mean	\$26,295.32	
98	Standard Deviation	\$29,842.83	

← df\$zscore <- (df\$cost -mean(df\$cost))

## Coefficient of Variation

- The coefficient of variation (CV) provides a relative measure of dispersion in data relative to the mean:

$$CV = \frac{\text{standard deviation}}{\text{mean}}$$

- Sometimes expressed as a percentage (x100)
- Provides a relative measure of risk to return
- Useful when comparing variability of two or more data sets with different scales
- Smaller CV → smaller risk
- Reciprocal of CV → return to risk

## Eg: Applying the Coefficient of Variation

- *Closing Stock Prices database*
- Which investment is **most** risky?
- Which investment would have **least** risk?

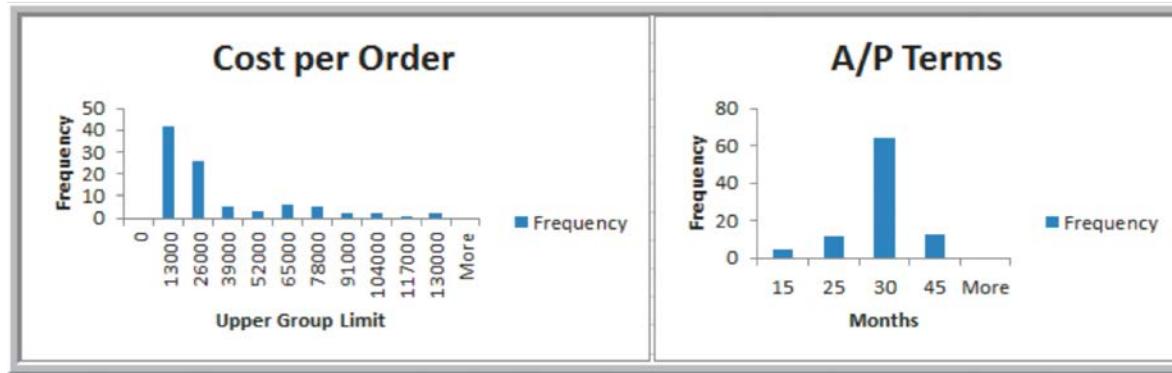
	A	B	C	D	E	F
1	<b>Closing Stock Prices</b>					
2						
3	<b>Date</b>	<b>IBM</b>	<b>INTC</b>	<b>CSCO</b>	<b>GE</b>	<b>DJ Industrials Index</b>
4	9/3/2010	\$127.58	\$18.43	\$21.04	\$15.39	10447.93
5	9/7/2010	\$125.95	\$18.12	\$20.58	\$15.44	10340.69
6	9/8/2010	\$126.08	\$17.90	\$20.64	\$15.70	10387.01
22	9/30/2010	\$134.14	\$19.20	\$21.90	\$16.25	10788.05
23	10/1/2010	\$135.64	\$19.32	\$21.91	\$16.36	10829.68
24	<b>Mean</b>	\$130.93	\$18.81	\$21.50	\$16.20	\$10,639.98
25	<b>Standard Deviation</b>	\$3.22	\$0.50	\$0.52	\$0.35	\$171.94
26	<b>Coefficient of Variation</b>	0.025	0.027	0.024	0.022	0.016

Intel (INTC) is **slightly riskier** than the other stocks.

Index fund has the **least risk** (lowest CV).

# Measures of Shape: Skewness

- **Skewness** describes the **lack of symmetry** of data.
  - Distributions that **tail off to the right** are called **positively skewed**; those that **tail off to the left** are said to be **negatively skewed**.



Positively skewed

Symmetrical

# Coefficient of Skewness

- Coefficient of Skewness (CS):

$$CS = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^3}{\sigma^3}$$

- ▶ CS is negative for left-skewed data.
- ▶ CS is positive for right-skewed data.
- ▶  $|CS| > 1$  suggests high degree of skewness.
- ▶  $0.5 \leq |CS| \leq 1$  suggests moderate skewness.
- ▶  $|CS| < 0.5$  suggests relative symmetry.

In R, can be obtained using the “skew” function in “psych” package.

# Eg: Measuring Skewness

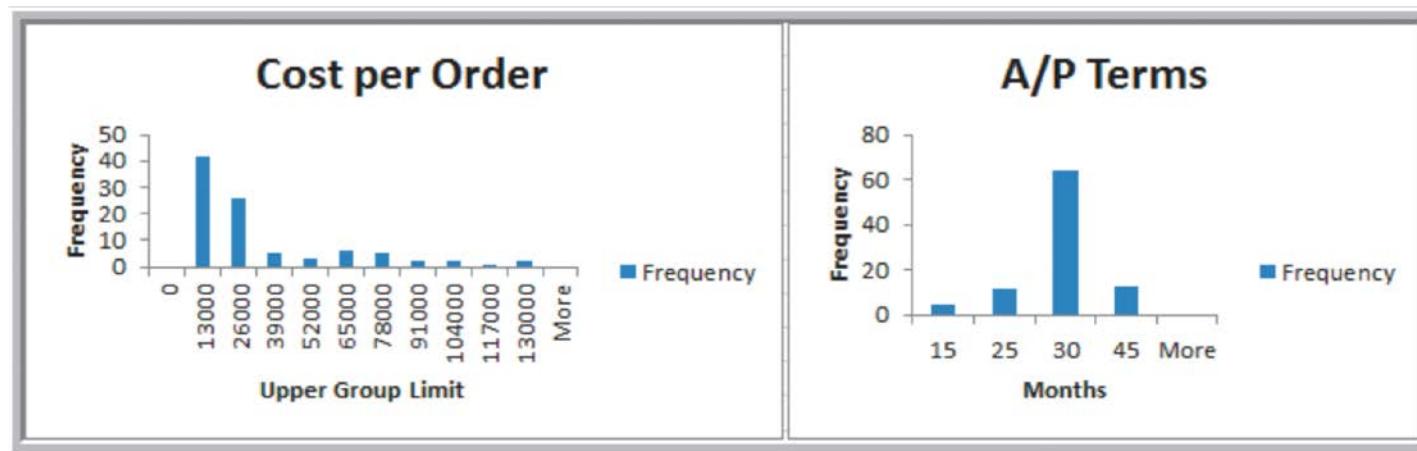
- Using *Purchase Orders* database

```
> skew(Cost)
```

```
[1] 1.611533
```

```
> skew(AP)
```

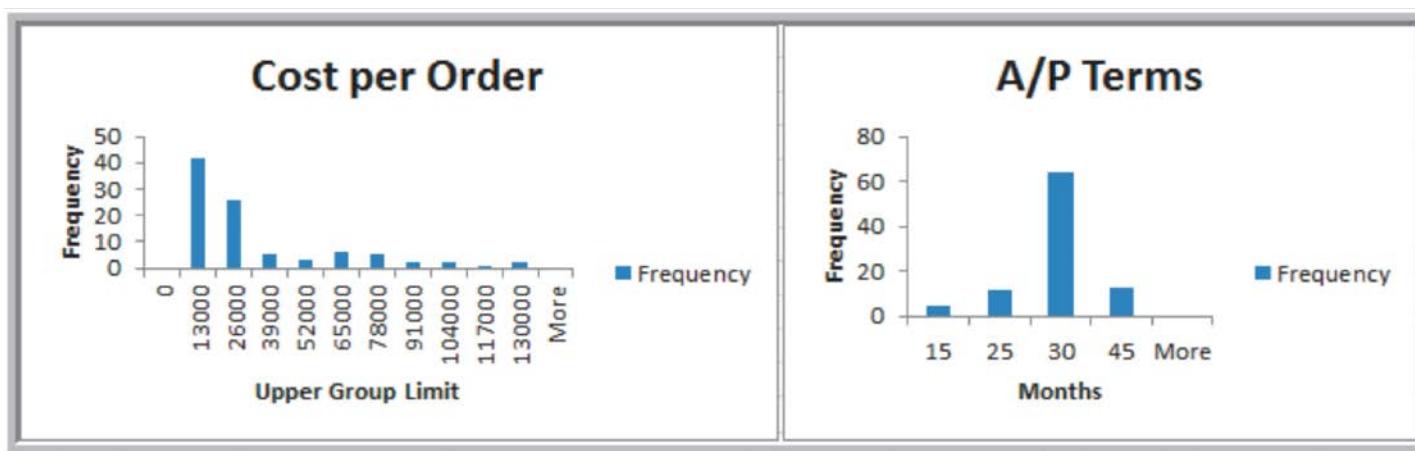
```
[1] 0.5802752
```



# Eg: Measuring Skewness

- Using *Purchase Orders* database
- Cost per order data:  $CS = 1.61$
- A/P terms data:  $CS = 0.58$

Which has higher skewness?  
Positive or Negative?



$CS = 1.61$   
High positive skewness

$CS = 0.58$   
Moderate positive skewness

# Shape and Measures of Location

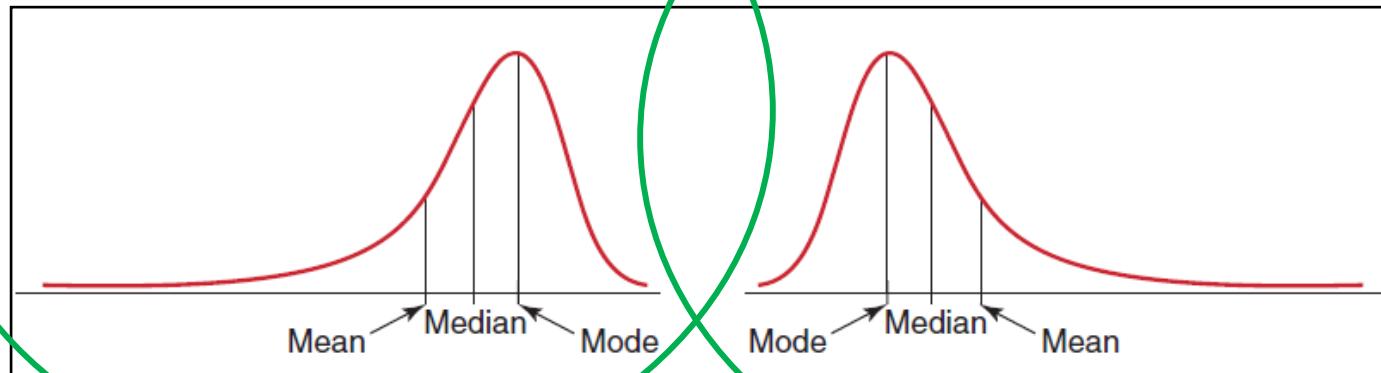
Comparing measures of location can sometimes reveal information about the shape of the distribution of observations.

Negatively skewed

Mean < Median < Mode

Positively skewed

Mode < Median < Mean



For example:

- If distribution was perfectly symmetrical and unimodal, the mean, median, and mode would all be the same.
- If it were negatively skewed, mean < median < mode
- Positive skewness would suggest that mode < median < mean

# Measures of Shape: Kurtosis

- Kurtosis refers to the **peakedness** (i.e., high, narrow) or **flatness** (i.e., short, flat-topped) of a histogram.
- Coefficient of kurtosis (CK) measures the **degree of kurtosis of a population**

$$CK = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^4}{\sigma^4}$$

- ▶ CK < 3 indicates the **data is somewhat flat with a wide degree of dispersion**.
- ▶ CK > 3 indicates the **data is somewhat peaked with less dispersion**.

In R, `kurtosi` function in the `psych` package can be used to compute CK (note however the cut off is 0 instead of 3)

```
> kurtosi(Cost)  
[1] 1.803636
```

```
> kurtosi(AP)  
[1] 1.277305
```

Kurtosis > 1, hence data is somewhat peaked with less dispersion

# Descriptive Statistics for Grouped Data

- Population mean:

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

- Sample mean:

---

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

- Population variance:

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

- Sample variance:

---

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

## Eg: Computing Statistical Measures from Frequency Distributions

- Computer Repair Times

A	B	C	D	E	F
1 Computer Repair Times					
3	Days (x)	Frequency (f)	Frequency*Days	Days - Mean	(Days - mean )^2
4	0	0	0	-14.912	222.368
5	1	0	0	-13.912	193.544
6	2	0	0	-12.912	166.720
7	3	0	0	-11.912	141.896
43	39	1	39	24.088	580.232
44	40	1	40	25.088	629.408
45	41	0	0	26.088	680.584
46	42	0	0	27.088	733.760
47	Sum	250	3728		8840.064
48					
49	Mean		14.912	Variance	35.50226506

```
> describe(x)
```

```
vars n mean sd median trimmed mad min max range skew kurtosis se
X1 1 250 14.91596 14 14.12 4.45 5 40 35 1.67 3.92 0.38
```

$$5.96^2 = 35.50$$

# Eg: Computing Home Value by Type and Region

function in  
'psych' package

```
> describeBy(df11$`Market Value`, group=df11$`Type`)
```

```
Descriptive statistics by group
group: A1
  vars n      mean      sd median trimmed      mad      min      max
X1    1 12 94166.67 10261.52 90500  93450 10452.33 81500 114000
      range skew kurtosis      se
X1 32500 0.54   -1.11 2962.25
-----
group: A2
  vars n      mean      sd median trimmed      mad      min      max range
X1    1 10 95000 8921.88 92300  93700 7042.35 87200 113200 26000
      skew kurtosis      se
X1 0.73   -0.91 2821.35
-----
group: B1
  vars n      mean      sd median trimmed      mad      min      max range
X1    1  8 88812.5 13569.23 85400 88812.5 4744.32 76600 120700 44100
      skew kurtosis      se
X1 1.49    0.91 4797.45
-----
group: B2
  vars n      mean      sd median trimmed      mad      min      max range
X1    1  5 89360 6209.91 91300  89360 3113.46 78800 94200 15400
      skew kurtosis      se
X1 -0.83   -1.22 2777.16
-----
group: C
  vars n      mean      sd median trimmed      mad      min      max
X1    1  7 89942.86 12625.22 87600 89942.86 9340.38 79800 116100
      range skew kurtosis      se
X1 36300 1.12   -0.25 4771.88
```

	House Age	Square Feet	Market Value	Type	Region	Sub-Reg
1	33	1812	90000	A1	1	U
2	32	1914	104400	A2	1	X
3	32	1842	93300	B2	1	Z
4	33	1812	91000	A1	1	U
5	32	1836	101900	A1	2	U
6	33	2028	108500	A1	2	U
7	32	1732	87600	A2	2	U
8	33	1850	96000	A2	2	U
9	32	1791	89200	B2	3	U
10	33	1666	88400	A1	3	U
11	32	1852	100800	A1	3	U
12	32	1620	96700	A1	3	U

## Descriptive Statistics for Categorical Data: The Proportion

- Proportion ( $p$ ), is the fraction of data that have a certain characteristic.
- Proportions are key descriptive statistics for categorical data, such as defects or errors in quality control applications or consumer preferences in market research.

# Eg: Computing a Proportion

- ▶ Proportion of orders placed by Spacetime Technologies

A	B	C	D	E	F	G	H	I	J
1 Purchase Orders									
2									
3 Supplier	Order No.	Item No.	Item Description	Item Cost	Quantity	Cost per order	A/P Terms (Months)	Order Date	Arrival Date
4 Spacetime Technologies	A0111	6489	O-Ring	\$ 3.00	900	\$ 2,700.00	25	10/10/11	10/18/11
5 Steelpin Inc.	A0115	5319	Shielded Cable/ft.	\$ 1.10	17,500	\$ 19,250.00	30	08/20/11	08/31/11
6 Steelpin Inc.	A0123	4312	Bolt-nut package	\$ 3.75	4,250	\$ 15,937.50	30	08/25/11	09/01/11
7 Steelpin Inc.	A0204	5319	Shielded Cable/ft.	\$ 1.10	16,500	\$ 18,150.00	30	09/15/11	10/05/11
8 Steelpin Inc.	A0205	5677	Side Panel	\$ 195.00	120	\$ 23,400.00	30	11/02/11	11/13/11
9 Steelpin Inc.	A0207	4312	Bolt-nut package	\$ 3.75	4,200	\$ 15,750.00	30	09/01/11	09/10/11
10 Alum Sheeting	A0223	4224	Bolt-nut package	\$ 3.95	4,500	\$ 17,775.00	30	10/15/11	10/20/11

$$\text{Proportion} = \frac{\text{Number of Orders by SpaceTime Technologies}}{\text{Total number of Orders}}$$

```
> df9<-Purchase_Orders
> length(df9$Supplier[df9$Supplier=="Spacetime Technologies"])
[1] 12
can also use filter function in dplyr
> nrow(df9)
[1] 94
> length(df9$Supplier[df9$Supplier=="Spacetime Technologies"])/nrow(df9)
[1] 0.1276596
```

# Measures of Association

- Data from 49 top liberal arts and research universities can be used to answer questions:
  - Is *Top 10% HS* related to *Graduation %*?
  - Is *Accept. Rate* related to *Expenditures/Student?*
  - Is *Median SAT* related to *Acceptance Rate?*

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	\$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	\$ 17,554	58	88
7	Berkeley	University	1176	37%	\$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	\$ 25,703	78	90
9	Brown	University	1281	24%	\$ 24,201	80	90
10	Bryn Mawr	Lib Arts	1255	56%	\$ 18,847	70	84

# Measures of Association - Covariance

- Covariance is a measure of the linear association between two variables,  $X$  and  $Y$ .

- For a population:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^N (x_i - \mu_x)(y_i - \mu_y)}{N}$$

- For a sample:

$$\text{cov}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Positive Covariance → direct relationship

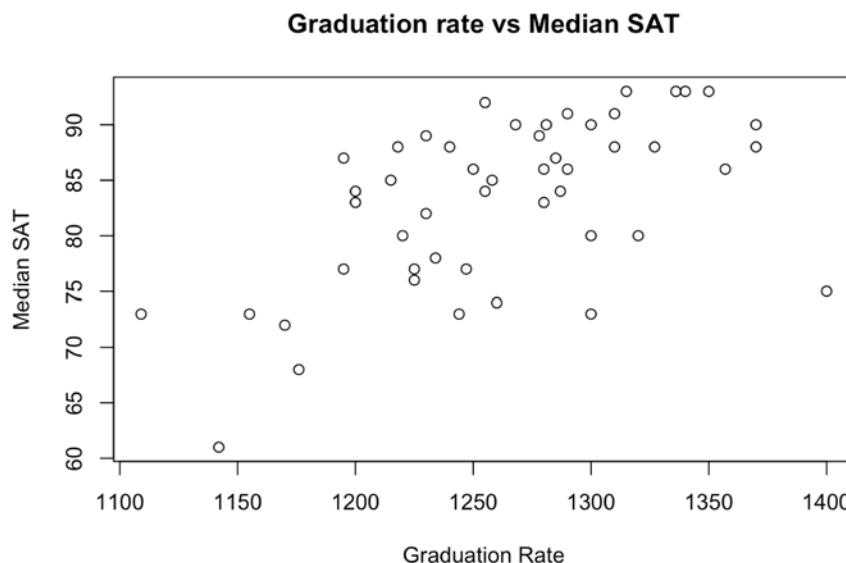
Negative Covariance → inverse relationship

Magnitude → degree of association

# Measures of Association - Covariance

Eg: Computing the Covariance

- Scatterplot of the *Colleges and Universities* data



```
> cov(df9$`Median SAT`, df9$`Graduation %`)
[1] 263.3703
```

# Measures of Association - Correlation

- Correlation is a measure of the linear association between two variables,  $X$  and  $Y$  (not dependent on units of measurement)

- Correlation Coefficient formulas: For a population:

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y}$$

For a sample:

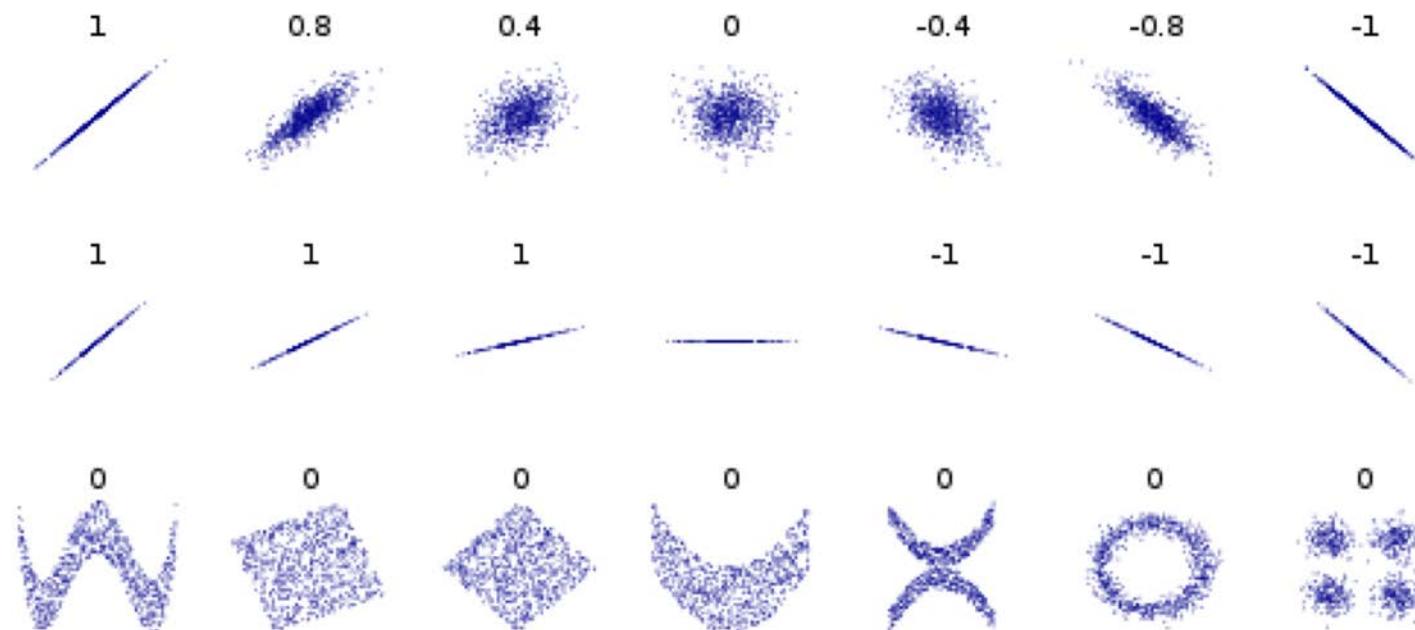
$$r_{xy} = \frac{\text{cov}(X, Y)}{s_x s_y}$$

- Range: -1 (Strong negative) and 1 (Strong positive linear relationship)
- 0 indicates no linear relationship
- Also known as: Pearson product moment correlation or Pearson's correlation coefficient

```
> cor(df9$`Median SAT`, df9$`Graduation %`)
[1] 0.5641468
```

# Measures of Association

- Correlation as a measure of LINEAR association



Source: Wikipedia (Correlation and dependence)

# Computing Correlation of Multiple Variables

```
> library(psych)
> cts <- corr.test(df9[3:7])
> cts
Call:corr.test(x = df9[3:7])
Correlation matrix
```

	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
Median SAT	1.00	-0.60	0.57	0.50	0.56
Acceptance Rate	-0.60	1.00	-0.28	-0.61	-0.55
Expenditures/Student	0.57	-0.28	1.00	0.51	0.04
Top 10% HS	0.50	-0.61	0.51	1.00	0.14
Graduation %	0.56	-0.55	0.04	0.14	1.00
Sample Size					

```
[1] 49
```

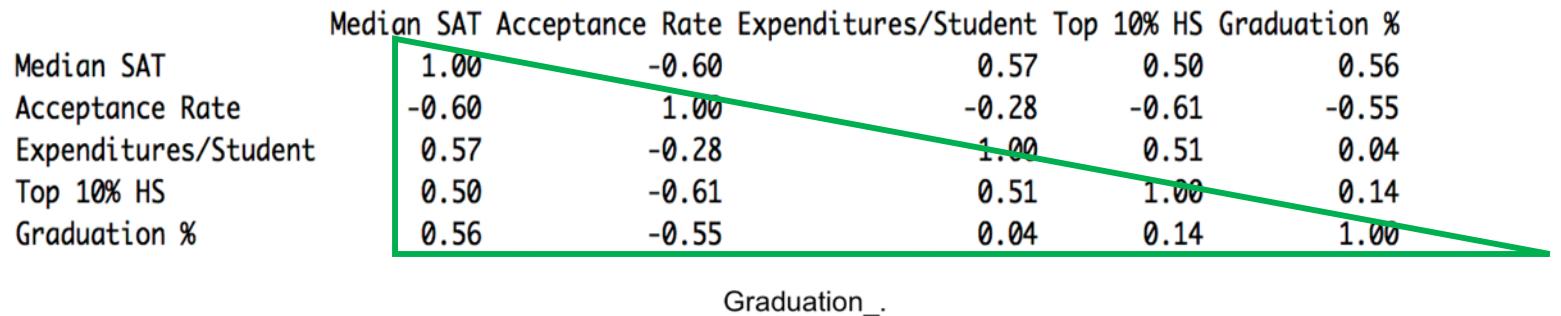
Probability values (Entries above the diagonal are adjusted for multiple tests.)

	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
Median SAT	0	0.00	0.00	0.00	0.00
Acceptance Rate	0	0.00	0.14	0.00	0.00
Expenditures/Student	0	0.05	0.00	0.00	0.77
Top 10% HS	0	0.00	0.00	0.00	0.68
Graduation %	0	0.00	0.77	0.34	0.00

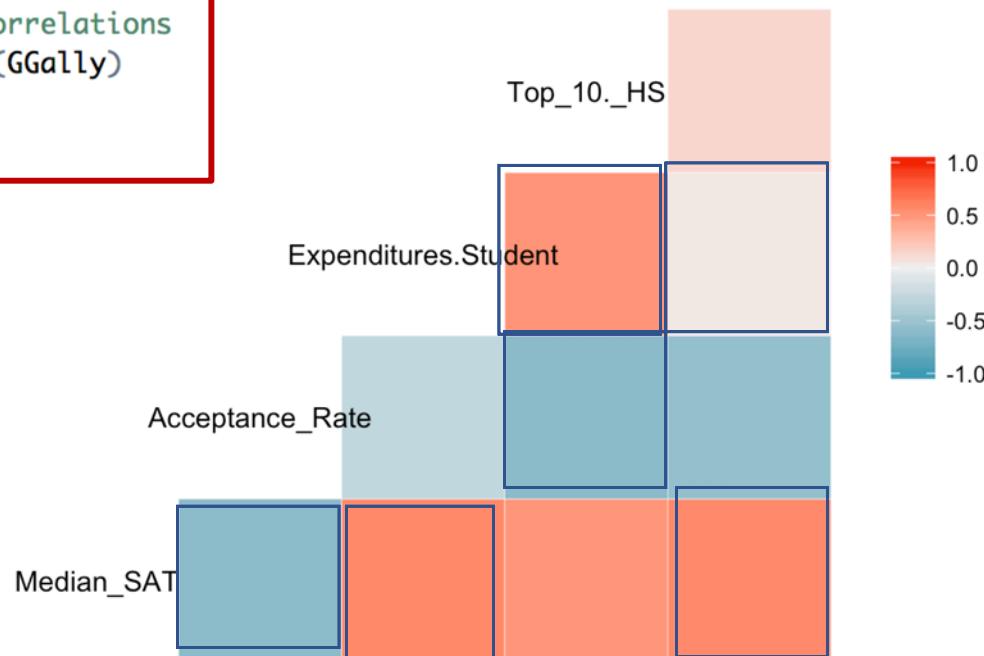
To see confidence intervals of the correlations, print with the short=FALSE option

.

# Plotting a correlation matrix



```
#plot graph of correlations  
install.packages(GGally)  
library(GGally)  
ggcorr(df9[3:7])
```



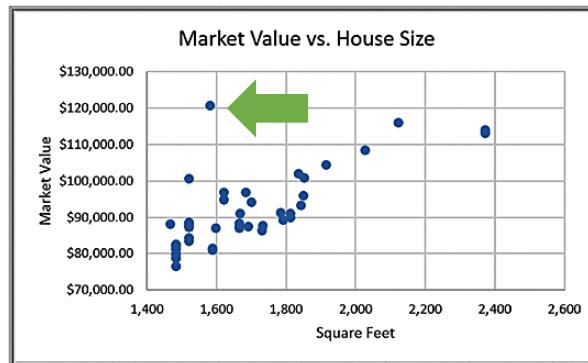
# Outliers

- Mean and Range are sensitive to outliers
- No standard definition of what constitutes an outlier.
- How do we identify potential outliers?
- Some rules of thumbs:
  - ▶ z-scores  $> +3$  or  $< -3$  ( $< 0.3\%$  for normal data)
  - ▶ Extreme outliers are  $> 3 \times \text{IQR}$  to the left of  $Q_1$  or right of  $Q_3$
  - ▶ Mild outliers are between  $(1.5 \text{ to } 3) \times \text{IQR}$  to the left of  $Q_1$  or right of  $Q_3$

# Eg: Investigating Outliers

- *Home Market Value* data

A	B	C	D	E	
1	Home Market Value				
2	House Age	Square Feet	z-score	Market Value	z-score
4	33	1,812	0.5300	\$90,000.00	-0.196
5	32	1,914	0.9931	\$104,400.00	1.168
6	32	1,842	0.6662	\$93,300.00	0.117
7	33	1,812	0.5300	\$91,000.00	-0.101
41	27	1,484	-0.9592	\$81,300.00	-1.020
42	27	1,520	-0.7957	\$100,700.00	0.818
43	28	1,520	-0.7957	\$87,200.00	-0.461
44	27	1,684	-0.0511	\$96,700.00	0.439
45	27	1,581	-0.5188	\$120,700.00	2.713
46	Mean	1,695		92,069	
47	Standard Deviation	220.257		10553.083	



- None of the z-scores exceed 3. However, while individual variables might not exhibit outliers, combinations of them might.
  - The last observation has a high market value (\$120,700) but a relatively small house size (1,581 square feet) and may be an outlier.

# What do you do with outliers?

- Leave them in the data if it is important
- Remove them if they are different from the rest
- Correct error in data entry

# Statistical Thinking in Business DM

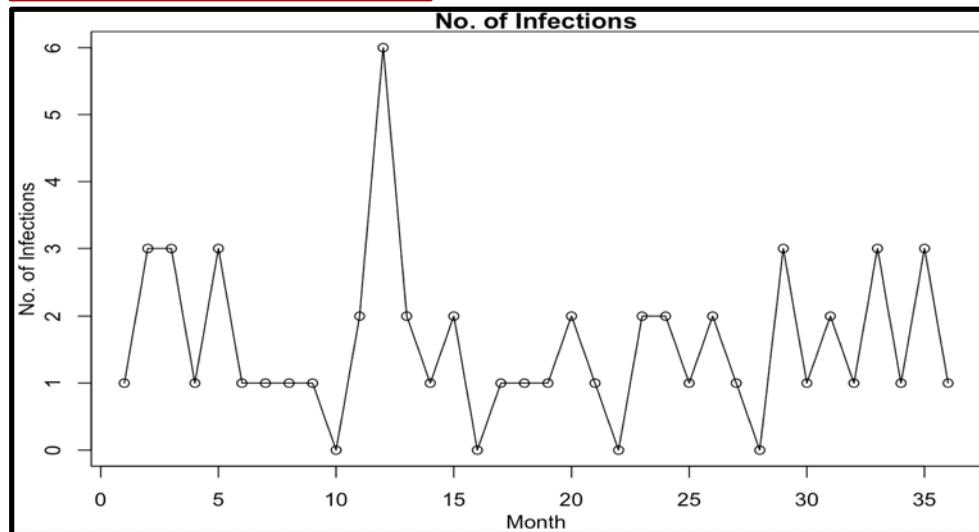
- Statistical Thinking is a philosophy of learning and action for improvement, based on principles that:
  - all work occurs in a system of interconnected processes
  - variation exists in all processes
  - better performance results from understanding and reducing variation
- Business Analytics provide managers with insights into facts and relationships that enables them to make better decisions.

# Applying Statistical Thinking

- Excel file *Surgery Infections*
  - Is month 12 simply random variation or some explainable phenomenon?

A	B
1	Surgery Infections
2	
3	Month Infections
4	1
5	2
6	3
7	4
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	12
16	13
17	14
18	15
19	16
20	17
21	18
22	19
23	20
24	21
25	22

```
> mean(df10$Infections)
[1] 1.583333
> sd(df10$Infections)
[1] 1.180194
```



```
> plot(df10$Month,df10$Infections, Main="No. of Infections",xlab="Month",ylab="No. of Infections")
> lines(df10$Month,df10$Infections)
```

# Applying Statistical Thinking

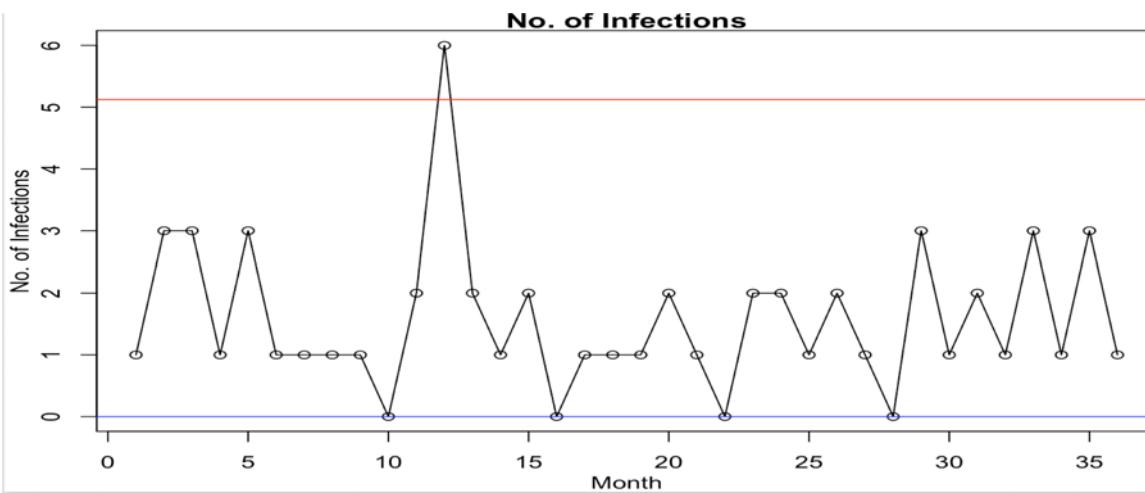
- Excel file *Surgery Infections*
  - Is month 12 simply random variation or some explainable phenomenon?

A	B
<b>Surgery Infections</b>	
1	
2	
3	<b>Month</b>
4	1
5	2
6	3
7	4
8	5
9	6
10	7
11	8
12	9
13	10
14	11
15	12
16	13
17	14
18	15
19	16
20	17
21	18
22	19
23	20
24	21
25	22

Applying the 3 std dev empirical rule!

```
> LL<- mean(df10$Infections)-3*(sd(df10$Infections))
> UL<- mean(df10$Infections)+3*(sd(df10$Infections))
> LL           > UL
[1] -1.957248 [1] 5.123914
```

```
abline(h=UL,col="red")
abline(h=0,col="blue") #infection rate cannot be less than 0
```

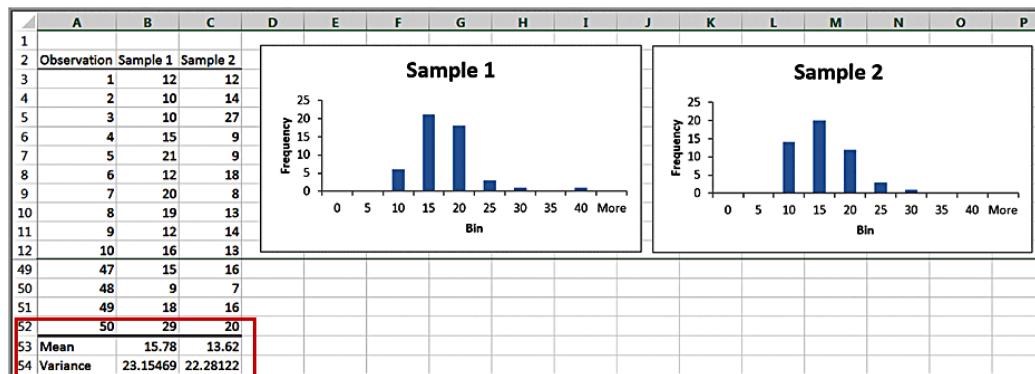


# Variability in Samples

- Different samples from any population will vary
  - different means, standard deviations, and other statistical measures
  - differences in shapes of histograms
- Samples are extremely sensitive to the sample size – the number of observations included in the samples.

## Eg: Variation in Sample Data

- Samples from *Computer Repair Times* data
- Population statistics:  $\mu = 14.91$  days,  $\sigma^2 = 35.5$  days<sup>2</sup>
- Two samples of size 50:



- Two samples of size 25:

