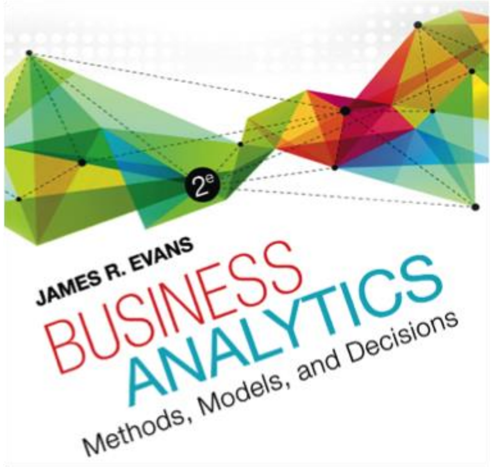


Chapter 4 Descriptive Statistical Measures

Adapted by Michael Hahsler

Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- 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

vs.

- **Sample**
  - a random subset of the population
  - a list of individuals who rented a comedy from Netflix in the past year

- **Statistics**
  Design a sample to **obtain sufficient information to draw a valid conclusion** about a population.

vs.

- **Data Science**
  - Often just the data we have to work with.
  - A subset of a dataset that is too large for our computer

Is the Netflix sample above a good sample? Why?
Other ways to select a sample?

Understanding Statistical Notation

- We typically label the elements of a data set using subscripted variables, x_1, x_2 , … , and so on, where x_i represents the ith observation. Upper-case letters like *X* represent often random variables.

- It is common practice in statistics to use
  - Greek letters, such as μ (mu; mean), σ (sigma; std. deviation), and π (pi; proportion), to represent population measures and
  - italic letters such as by $\bar{x}$ (called *x*-bar), *s*, and *p* to represent sample statistics.

- *N* represents the number of items in a population and *n* represents the number of observations in a sample.

Content

- Notation
- Measures of Location
  - Mean
  - Median
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers

Measures of Location: Arithmetic Mean

- Sample mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

- Excel function: =AVERAGE(*data range*)

- **Outliers can affect the value of the mean.**
- **Mean valid for interval/ratio variables and often questionable for ordinal variables.**

Outliers

Person	Age
1	17
2	21
3	15
4	18
5	999
6	22
7	11
8	25
Mean	141.00

Person	Age
1	17
2	21
3	15
4	18
5	
6	22
7	11
8	25
Mean	18.43

Wikipedia: In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to **variability in the measurement** or it may indicate **experimental error**; the latter are sometimes excluded from the data set.

Measures of Location: Median

- The **median** specifies the middle value when the data are arranged from least to greatest.
  - Half the data are below the median, and half the data are above it.
  - For an odd number of observations, the median is the middle of the sorted numbers.
  - For an even number of observations, the median is the mean of the two middle numbers.
- We could use the Sort option in Excel to rank-order the data and then determine the median. The Excel function =MEDIAN(*data range*) could also be used.

- **The median is meaningful for ratio, interval, and ordinal data.**
- **Not affected by outliers.**

Median

Person	Age
1	17.00
2	21.00
3	15.00
4	18.00
5	999.00
6	22.00
7	11.00
8	25.00
Mean	141.00
Median	19.50

Median is insensitive to outliers!

Using Measures of Location – Example 4.5: Quoting Computer Repair Times

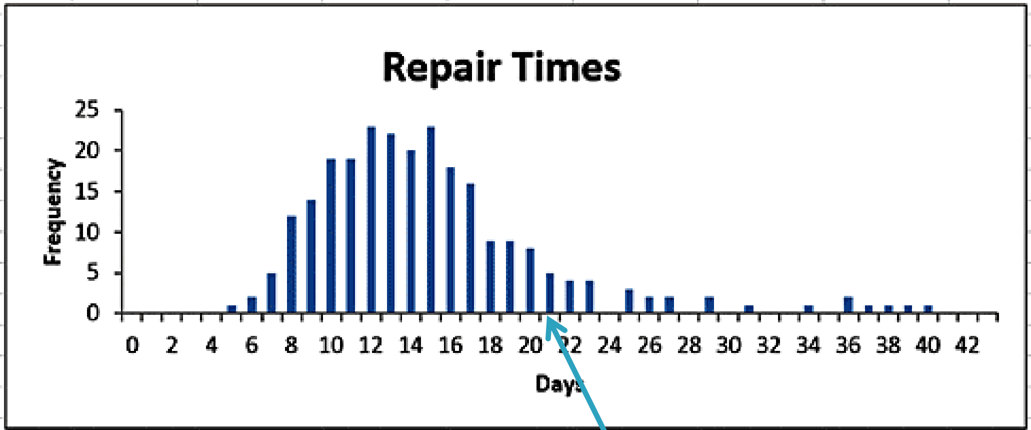
The Excel file *Computer Repair Times* includes 250 repair times for customers.

- What repair time would be reasonable to quote to a new customer?
- Median repair time is 2 weeks; mean and mode are about 15 days.
- Examine the histogram.

	A	B
1	Computer Repair Times	
2		
3	Sample	Repair Time (Days)
4	1	18
5	2	15
6	3	17
250	247	31
251	248	6
252	249	17
253	250	13
254		
255	Mean	14.912
256	Median	14
257	Mode	15

Example 4.5 (continued)

Computer Repair Times			
		Relative	Cumulative
Days	Frequency	Frequency	Percentage
0	0	0.000	0.0%
1	0	0.000	0.0%
2	0	0.000	0.0%
3	0	0.000	0.0%
4	0	0.000	0.0%
5	1	0.004	0.4%
6	2	0.008	1.2%
7	5	0.020	3.2%
8	12	0.048	8.0%
9	14	0.056	13.6%
10	19	0.076	21.2%
11	19	0.076	28.8%
12	23	0.092	38.0%
13	22	0.088	46.8%



90% are completed within 3 weeks

Distribution is important!

Content

- Notation
- Measures of Location
- Measures of Dispersion
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Empirical Rules
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers

Measures of Dispersion: Range

- The **range** is the simplest and is the difference between the maximum value and the minimum value in the data set.
- In Excel, compute as =MAX(*data range*) - MIN(*data range*).
- The range is **affected by outliers** and is often used only for very small data sets.

Measures of Dispersion: Interquartile Range

- The **interquartile range (IQR)**, or the **midspread** is the difference between the first and third quartiles, Q3 – Q1.

- This includes only the middle 50% of the data and, therefore, is **not influenced by extreme values**.

Example Purchase Orders data

- For the Cost per order data:
  - Third Quartile = Q_3 = $27,593.75
  - First Quartile = Q_1 = $6,757.81
- Interquartile Range = $27,593.75 – $6,757.81 =$20,835.94

Measures of Dispersion: Variance

- The variance is the "average" of the squared deviations from the mean.
- For a population:

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

  - In Excel: =VAR.P(*data range*)
- For a sample:

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

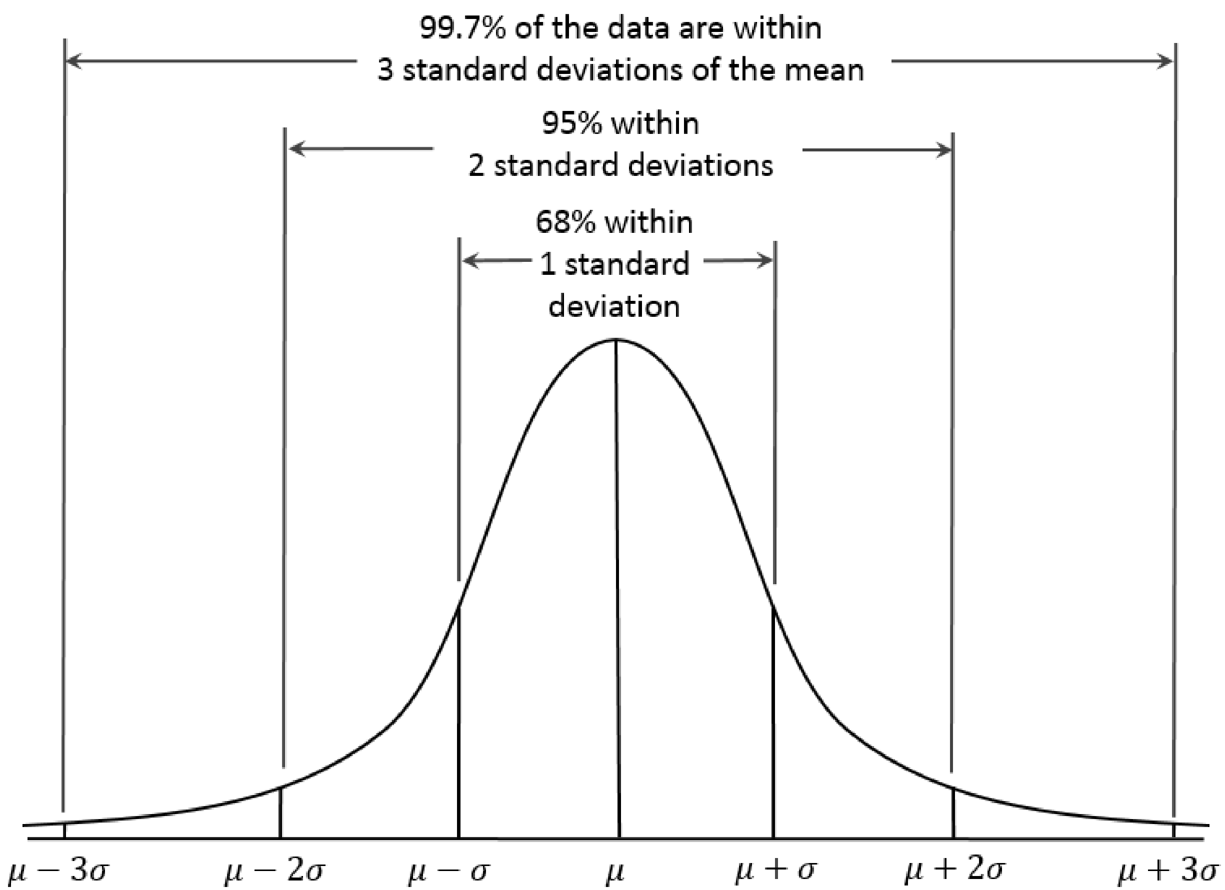
  - In Excel: =VAR.S(*data range*)

- Note the difference in denominators!

- The **standard deviation** is the square root of the variance.

Empirical Rules

- The empirical Rule comes from the normal distribution.



Caution: Most data does not follow a normal distribution!

Chebyshev's Theorem

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

- Examples:
  - For k = 2: at least ¾ or 75% of the data lie within two standard deviations of the mean
  - For k = 3: **at least 8/9 or 89% of the data lie within three standard deviations of the mean**

Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers

Standardized Values

- A **standardized value**, commonly called a ***z*-score**, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement.
- The *z*-score for the i^{th} observation in a data set is calculated as follows:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{4.9}$$

  - Excel function: =STANDARDIZE(*x, mean, standard_dev*).

Standardized data is needed by many predictive methods since it makes variables comparable.

Example 4.12 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	0
98	Standard Deviation	$29,842.83	1

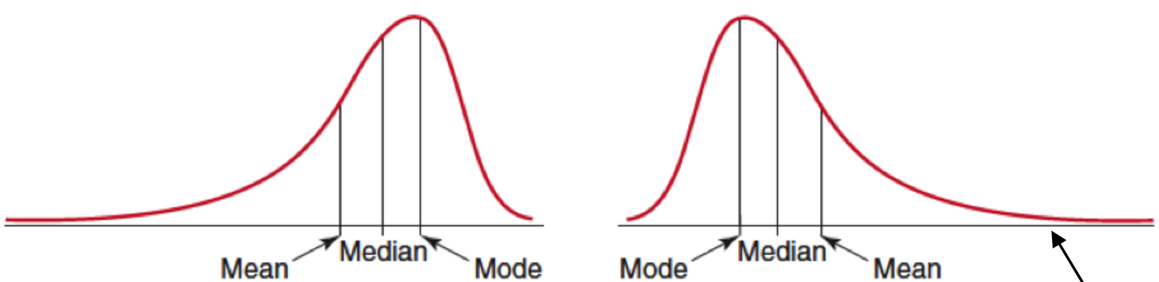
=(B2 - B97)/B98, or
=STANDARDIZE(B2,B97,B98).

Shape and Measures of Location

- Comparing measures of location can sometimes reveal information about the shape of the distribution of observations.
  - For example, if the distribution were **perfectly symmetrical** and unimodal, the mean, median, and mode would all be the same.
  - If it were **negatively skewed**, we would generally find that mean < median < mode
  - **Positive skewness** would suggest that mode < median < mean

Neg. skewed

pos. skewed



Long tails are very common

Content

Descriptive Statistics for Categorical Data: The Proportion

- The **proportion**, denoted by *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.
- Example: Proportion of female students is 60%.
- Example: Proportion of orders placed by Spacetime Technologies

=COUNTIF(A4:A97, “Spacetime Technologies”)/94

= 12/94 = 0.128

	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

Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
  - Correlation
- Outliers

Measures of Association

- Two variables have a strong statistical relationship with one another if they appear to “move” together.
- When two variables appear to be related, you might suspect a cause-and-effect relationship.
- **Caution: Correlation does not prove causation!** Statistical relationships may exist even though a change in one variable is not caused by a change in the other.

Measures of Association: Correlation

- **Correlation** is a measure of the linear relationship between two variables, *X* and *Y*, which does not depend on the units of measurement.
- Correlation is measured by the correlation coefficient, also known as the **Pearson product moment correlation coefficient**.
- Correlation coefficient for a population:

$$\rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \qquad (4.19)$$

- Correlation coefficient for a sample:

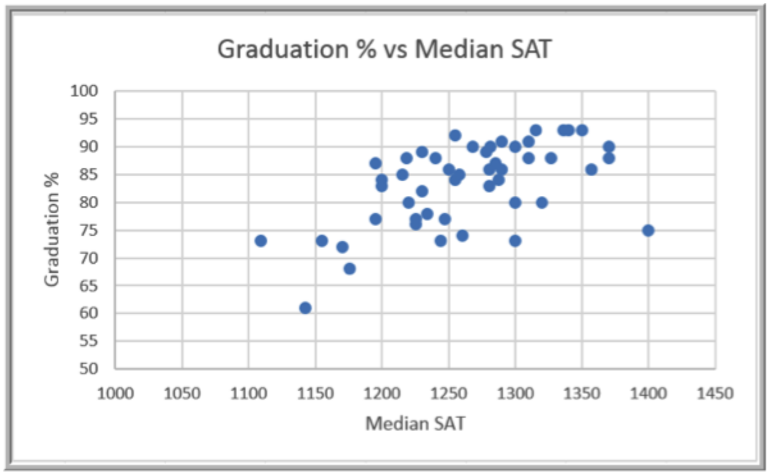
$$r_{xy} = \frac{\text{cov}(X, Y)}{s_x s_y} \qquad (4.20)$$

- The correlation coefficient is scaled between -1 and 1.
- Excel function: =CORREL(*array1*,*array2*)

Example 4.21

Computing the Correlation Coefficient

- *Colleges and Universities* data

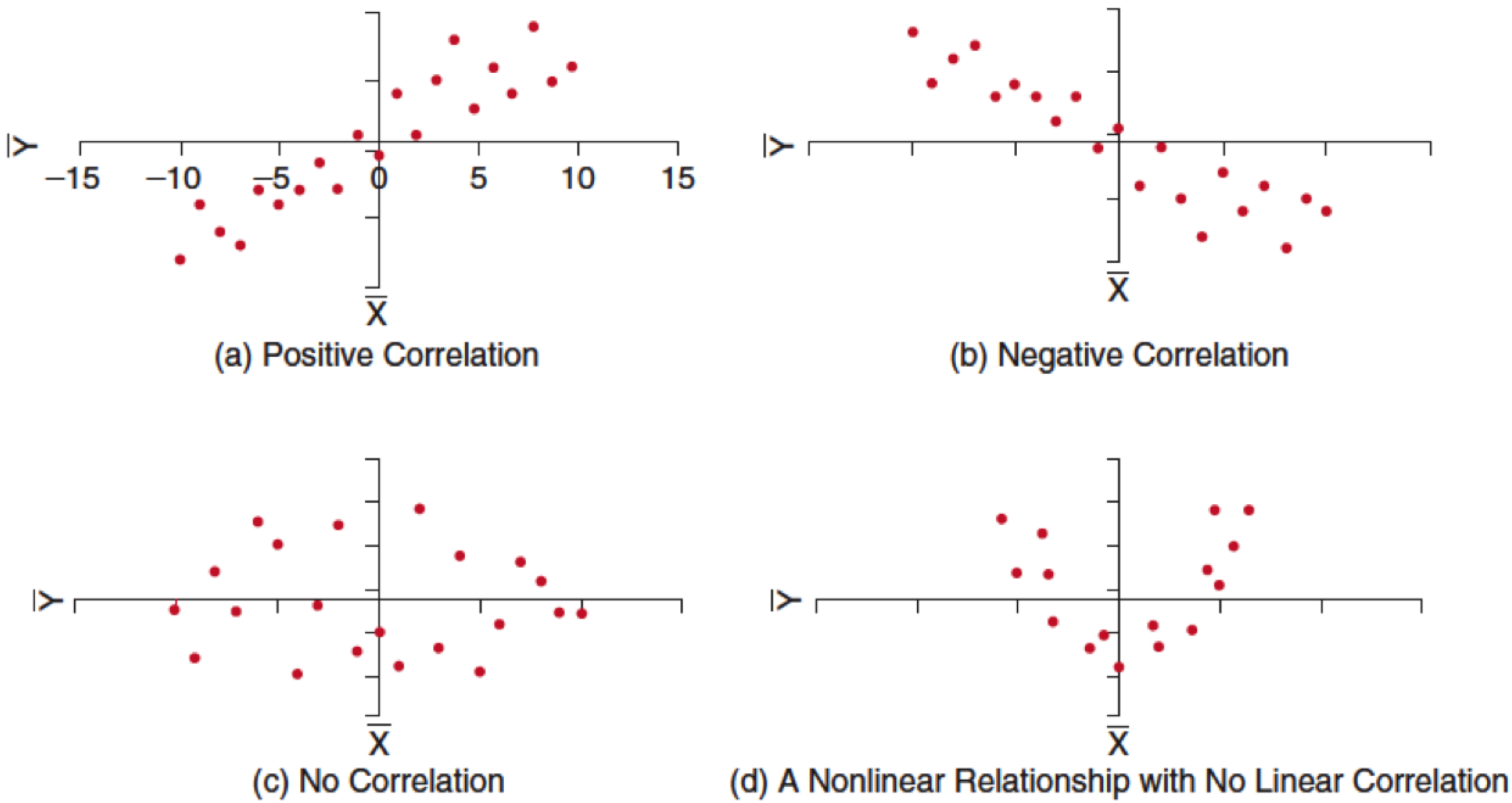


	A	B	C	D	E	F
1		Graduation % (X)	Median SAT (Y)	X - Mean(X)	Y - Mean(Y)	(X - Mean(X))(Y-Mean(Y))
2		93	1315	9.755	51.898	506.2698875
3		80	1220	-3.245	-43.102	139.8617243
4		88	1240	4.755	-23.102	-109.8525614
47		86	1250	2.755	-13.102	-36.09745939
48		91	1290	7.755	26.898	208.5964182
49		93	1336	9.755	72.898	711.1270304
50		93	1350	9.755	86.898	847.698459
51	Mean	83.245	1263.102		Sum	12641.77551
52	Standard Deviation	7.449	62.676		Count	49
53					Covariance	263.3703231
54					Correlation	0.564146827
55						
56					CORREL Function	0.564146827

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

Is there a causal relationship?

Examples of Correlation



Why is correlation important?

Association between categorical and continuous Variables

Group by the categorical variable and aggregate using

- Average
- Max and Min
- Product
- Standard deviation
- Variance

This is a PivotTable!

Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers

Identifying Outliers

- **There is no standard definition of what constitutes an outlier!**

- **Wikipedia: "***In statistics, an outlier is an observation point that is* ***distant from other observations****. […] Outliers can occur by chance in any distribution, but they often indicate either* ***measurement error*** *or that the population has a* ***heavy-tailed distribution****."*

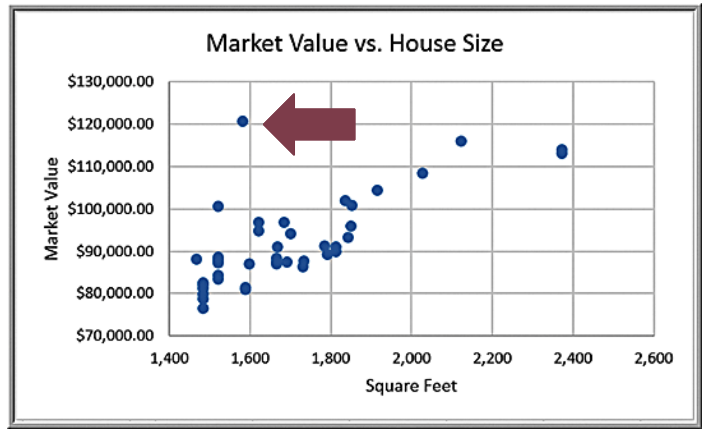
- If the outlier is due to a measurement error then we often want to exclude it from the analysis.

- Some typical rules of thumb:
  - **Look at histogram!**
  - **Normal distribution:** *z*-scores greater than +3 or less than -3
  - **Boxplot:**
    - Extreme outliers are more than 3*IQR to the left of Q_1 or right of Q_3
    - Mild outliers are between 1.5*IQR and 3*IQR to the left of Q_1 or right of Q_3

Example 4.23: Investigating Outliers

- *Home Market Value* data

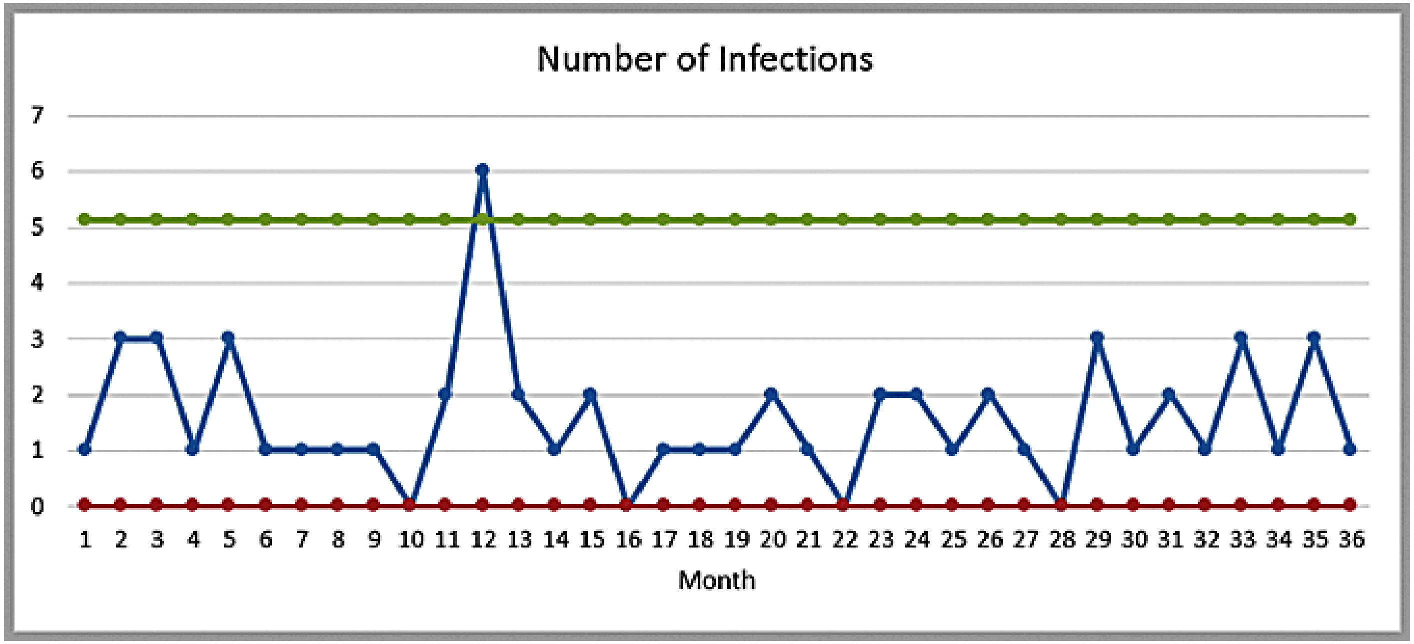
	A	B	C	D	E
1	Home Market Value				
2					
3	House Age	Square Feet	z-score	Market Value	z-score
4	33	1,812	0.5300	$90,000.00	-0.196
5	32	1,914	0.9931	$104,400.00	1.168
6	32	1,842	0.6662	$93,300.00	0.117
7	33	1,812	0.5300	$91,000.00	-0.101
41	27	1,484	-0.9592	$81,300.00	-1.020
42	27	1,520	-0.7957	$100,700.00	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.

Example 4.24 Outlier in a Medical Timeseries

- Three-standard deviation empirical rule:



- There is only a 0.3% (for normally distributed data) or a 11% (for any distribution) chance to see an observation outside +/- 3 std.dev.
- This suggests that month 12 is statistically different from the rest of the data.