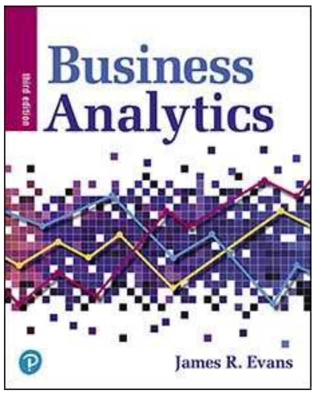
Business Analytics: Methods, Models, and Decisions

Third Edition



Chapter 4

Descriptive Statistics

 Slide - 1

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 individuals who rented a comedy from Netflix in the past year
- The purpose of sampling is to obtain sufficient information to draw a valid inference about a population.

 Slide - 2

Understanding Statistical Notation

- We typically label the elements of a data set using subscripted variables, $x_1, x_2, \ldots,$ and so on, where x_i represents the i^th observation.
- It is common practice in statistics to use Greek letters, such as μ(mu), σ(sigma), and π(pi), to represent population measures and italic letters such as by $\overline{x}$ (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.
- Capital Greek sigma represents summation: $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$

Pearson Slide - 3

Measures of Location: Arithmetic Mean

- Population mean: $$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \quad (4.4)$$
- Sample mean: $$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad (4.5)$$
- Excel function: = AVERAGE(*data range*)
- Property of the mean:

$$\sum_i (x_i - \overline{x}) = 0 \quad (4.6)$$

- Outliers can affect the value of the mean.

Pearson Slide - 4

Example 4.14: Computing Mean Cost Per Order

Purchase Orders database

- Using formula (4.5):

$$= \frac{\text{SUM(B2 : B95)}}{\text{COUNT(B2 : B95)}}$$

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

	A	B
1	Observation	Cost per order
2	x1	$82,875.00
3	x2	$42,500.00
4	x3	$24,150.00
5	x4	$22,575.00
6	x5	$19,250.00
91	x90	$467.50
92	x91	$9,975.00
93	x92	$30,625.00
94	x93	$21,450.00
95	x94	$68.75
96	Sum of Cost per Order	$2,471,760.00
97	Number of observations	94
98		
99	Mean Cost per Order (=B96/B97)	$26,295.32
100		
101	Excel AVERAGE function	$26,295.32

Using Excel AVERAGE Function

=AVERAGE(B2:B95)

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.

Example 4.15: Finding the Median Cost Per Order

- Sort the data from smallest to largest. Since we have 90 observations, the median is the average of the 47^{th} and 48^{th} observation.

$$\text{Median} = \frac{(\$15,562.50 + \$15,750.00)}{2}$$
$$= \$15,656.25$$

=MEDIAN(B2:B94)

	A	B	C	D
1	Rank	Cost per order		
2	1	\$68.75		
3	2	\$82.50		
4	3	\$375.00		
5	4	\$467.50		
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		
52	51	\$16,330.00		

Pearson Slide - 7

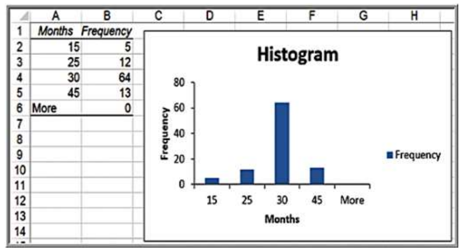
Measures of Location: Mode

- The **mode** is the observation that occurs most frequently.
- The mode is most useful for data sets that contain a relatively small number of unique values.
- You can easily identify the mode from a frequency distribution by identifying the value or group having the largest frequency or from a histogram by identifying the highest bar.
- Excel function: $= \text{MODE.SNGL}(\textit{data range})$.
- For multiple modes: $= \text{MODE.MULT}(\textit{data range})$

Pearson Slide - 8

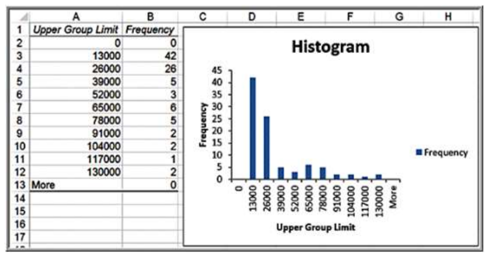
Example 4.16: Finding the Mode

- Purchase Orders database: A/P Terms
  - Mode = 30 months



Months	Frequency
15	5
25	12
30	64
45	13
More	0

- Cost per order
  - Mode is the group between $0 and $13,000.



Upper Group Limit	Frequency
0	0
13000	42
26000	26
39000	5
52000	3
65000	6
78000	5
91000	2
104000	2
117000	1
130000	2
More	0

Measures of Location: Midrange

- The **midrange** is the average of the greatest and least values in the data set.
- Caution must be exercised when using the midrange because extreme values easily distort the result. This is because the midrange uses only two pieces of data, whereas the mean uses all the data; thus, it is usually a much rougher estimate than the mean and is often used for only small sample sizes.

Example 4.17: Computing the Midrange

- *Purchase Orders* data
- Use the Excel MIN and MAX functions or sort the data and find them easily.
- Cost per order midrange:

=($68.78 + $1127,500)/2 = $63,784.39

 Slide - 11

Example 4.18: 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

 Slide - 12

Example 4.18 Continued

Computer Repair Times

Days	Frequency	Relative Frequency	Cumulative 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

Measures of Dispersion

- **Dispersion** refers to the degree of variation in the data; that is, the numerical spread (or compactness) of the data.
- Key measures:
  - Range
  - Interquartile range
  - Variance
  - Standard deviation

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.

Example 4.19: Computing the Range

- *Purchase Orders* data
- For the cost per order data:
  - Maximum = $127,500
  - Minimum = $68.78

Range = $127,500 − $68.78 = $127,431.22

Measures of Dispersion: Interquartile Range

- The **interquartile range (IQR)**, or the **midspread** is the difference between the first and third quartiles, $Q_3 - Q_1$.
- This includes only the middle 50% of the data and, therefore, is not influenced by extreme values.

Example 4.20: Computing the Interquartile Range

- *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:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \quad (4.7)$$

  - In Excel: $=\mathrm{VAR.P}(\mathit{data\ range})$
- For a sample:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} \quad (4.8)$$

  - In Excel: $=\mathrm{VAR.S}(\mathit{data\ range})$
- Note the difference in denominators!

 Slide - 19

Example 4.21: Computing the Variance

- Purchase Orders Cost per order data

	A	B	C	D
1	Observation	Cost per order	(xi - mean)	(xi - mean)^2
2	x1	$82,875.00	$56,579.68	$3,201,260,285.21
3	x2	$42,500.00	$16,204.68	$262,591,681.48
4	x3	$24,150.00	-$2,145.32	$4,602,394.25
5	x4	$22,575.00	-$3,720.32	$13,840,774.57
6	x5	$19,250.00	-$7,045.32	$49,636,521.91
91	x90	$467.50	-$25,827.82	$667,076,241.99
92	x91	$9,975.00	-$16,320.32	$266,352,817.12
93	x92	$30,625.00	$4,329.68	$18,746,136.27
94	x93	$21,450.00	-$4,845.32	$23,477,117.66
95	x94	$68.75	-$26,226.57	$687,832,929.32
96	Sum of Cost per Order	$2,471,760.00	Sum of squared deviations	$82,825,295,365.68
97	Number of observations	94		
98				
99	Mean Cost per Order (=B96/B97)	$26,295.32	Variance (=D96/(B97-1))	$890,594,573.82
100				
101			Excel VAR.S function	$890,594,573.82

 Slide - 20

Measures of Dispersion: Standard Deviation

- The **standard deviation** is the square root of the variance.
  - Note that the dimension of the variance is the square of the dimension of the observations, whereas the dimension of the standard deviation is the same as the data. This makes the standard deviation more practical to use in applications.
- For a population:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad (4.9)$$

  - In Excel: $= \text{STDEV.P}(\textit{data range})$
- For a sample:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (4.10)$$

  - In Excel: $= \text{STDEV.S}(\textit{data range})$

Example 4.22: Computing the Standard Deviation

- *Purchase Orders* Cost per order data
- Using the results of Example 4.21, take the square root of the variance:

$$\sqrt{890{,}594{,}573.82} = \$29{,}842.8312.$$

- Alternatively, use the Excel function =STDEV.S(B2:B95)

Standard Deviation as a Measure of Risk

Excel file: *Closing Stock Prices*

Intel (INTC):
Mean = $18.81
Standard deviation = $0.50

General Electric (GE):
Mean = $16.19
Standard deviation = $0.35

INTC is a higher risk investment than GE.

	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

Pearson Slide - 23

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 $1 - \frac{1}{k^2}$.
- Examples:
  - 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

Pearson Slide - 24

Example 4.23: Applying Chebyshev's Theorem

- *Purchase Orders* database
- A two-standard-deviation interval around the mean is [-$33,390.34, $85,980.98].
  - 89 of 94, or 94.7%, of the observations fall in this interval.
- A three-standard-deviation interval is [-$63,233.17, $115,823.81]
  - 92 of 94, or 97.9%, fall in this interval.

Empirical Rules

- For many data sets encountered in practice:
  - Approximately 68% of the observations fall within one standard deviation of the mean $\bar{x} - s$ and $\bar{x} + s$.
  - Approximately 95% fall within two standard deviations of the mean $\bar{x} \pm 2s$.
  - Approximately 99.7% fall within three standard deviations of the mean $\bar{x} \pm 3s$.
- These rules are commonly used to characterize the natural variation in manufacturing processes and other business phenomena.

Process Capability Index

- The 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:

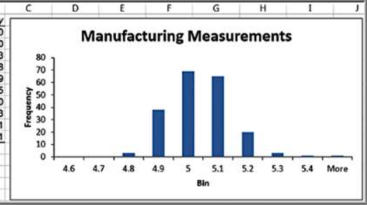
$$C_p = \frac{\text{Upper Specification} - \text{Lower Specification}}{\text{Total Variation}} \qquad (4.11)$$

Example 4.24: Using Empirical Rules to Measure the Capability of a Manufacturing Process

	A	B	C	D	E	F	G	H	I	J
1	Manufacturing Measurements									
2										
3	5.21	5.87	4.85	4.95	5.07	4.96	4.96	5.11	Mean	4.99
4	5.02	5.33	4.82	4.86	4.82	4.96	5.06	5.11	Standard deviation	0.117
5	4.90	5.11	5.02	5.13	5.03	4.94	4.86	5.08		
6	5.00	5.07	4.90	4.95	4.85	5.19	4.96	5.03	Mean - 3*Stdev	4.640
7	5.16	4.93	4.73	5.22	4.89	4.91	4.99	4.94	Mean + 3*Stdev	5.340
8	5.03	4.99	5.04	4.81	4.82	5.01	4.94	4.88	Total variaton	0.700
9	4.96	5.04	5.07	4.91	5.18	4.93	5.06	4.91		
10	5.04	5.14	4.81	4.95	5.02	5.05	4.95	4.86	Lower Specification	4.8
11	4.98	5.09	5.04	4.94	5.05	4.96	5.02	4.89	Upper Specification	5.2
12	5.07	5.06	5.03	4.81	4.88	4.92	5.01	4.91	Specification range	0.4
13	5.02	4.85	5.01	5.11	5.08	4.95	5.04	4.87		
14	5.08	4.93	5.14	4.81	4.98	5.08	5.01	4.93	Cp	0.57

Empirical rules

	A	B
1	Bin	Frequency
2	4.6	0
3	4.7	0
4	4.8	3
5	4.9	38
6	5	69
7	5.1	65
8	5.2	20
9	5.3	3
10	5.4	1
11	More	1



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:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{4.12}$$

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

Properties of *z*-Scores

$$z_i = \frac{x_i - \bar{x}}{s} \tag{4.12}$$

- The numerator represents the distance that x_i is from the sample mean; a negative value indicates that x_i lies to the left of the mean, and a positive value indicates that it lies to the right of the mean. By dividing by the standard deviation, *s*, we scale the distance from the mean to express it in units of standard deviations. Thus,
  - a *z*-score of 1.0 means that the observation is one standard deviation to the right of the mean;
  - a *z*-score of -1.5 means that the observation is 1.5 standard deviations to the left of the mean.

Example 4.25: Computing *z*-Scores

- *Purchase Orders* Cost per order data

	A	B	C	D
1	Observation	Cost per order	z-score	
2	x1	$82,875.00	1.90	=(B2-B96)/B97
3	x2	$42,500.00	0.54	
4	x3	$24,150.00	-0.07	
5	x4	$22,575.00	-0.12	
6	x5	$19,250.00	-0.24	
91	x90	$467.50	-0.87	
92	x91	$9,975.00	-0.55	
93	x92	$30,625.00	0.15	
94	x93	$21,450.00	-0.16	
95	x94	$68.75	-0.88	
96	Mean	$26,295.32		
97	Standard Deviation	$29,842.83		

Coefficient of Variation

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

$$\text{CV} = \frac{\text{Standard Deviation}}{\text{Mean}} \qquad (4.13)$$

  - Sometimes expressed as a percentage.
  - Provides a relative measure of risk to return.
- **Return to risk** = $\frac{1}{\text{CV}}$, is often easier to interpret, especially in financial risk analysis.
  - The Sharpe ratio is a related measure in finance.

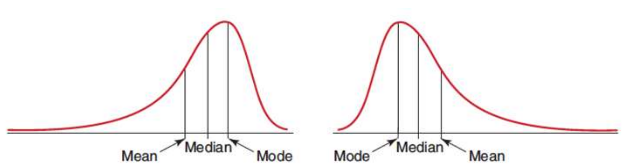
Example 4.26: Applying the Coefficient of Variation

- *Closing Stock Prices* worksheet
- Intel (INTC) is slightly riskier than the other stocks.
- The Index fund has the least risk (lowest CV).

	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
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	Mean	$130.93	$18.81	$21.50	$16.20	$10,639.98
25	Standard Deviation	$3.22	$0.50	$0.52	$0.35	$171.94
26	Coefficient of Variation	0.025	0.027	0.024	0.022	0.016

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.



Negatively skewed   Positively skewed

Coefficient of Skewness

- Coefficient of Skewness (CS):

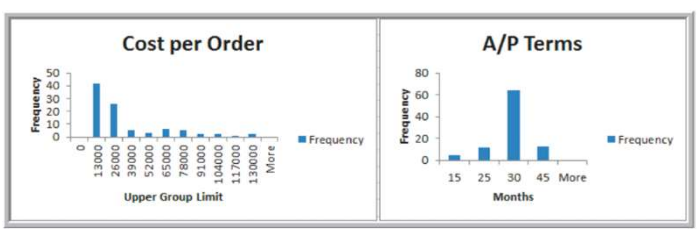
$$\mathrm{CS} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3} \qquad (4.14)$$

- Excel function: =SKEW(*data range*)
  - CS is negative for left-skewed data.
  - CS is positive for right-skewed data.
  - $|\mathrm{CS}| > 1$ suggests high degree of skewness.
  - $0.5 \le |\mathrm{CS}| \le 1$ suggests moderate skewness.
  - $|\mathrm{CS}| < 0.5$ suggests relative symmetry.

 Slide - 35

Example 4.27: Measuring Skewness

- *Purchase Orders* database
- Cost per order data: CS = 1.66 (high positive skewness)
- A/P terms data: CS = 0.60 (more symmetric)



 Slide - 36

Measures of Shape: Kurtosis

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

$$\mathrm{CK} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4} \qquad (4.15)$$

  - 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.

 Slide - 37

Excel Function for Kurtosis

- Excel computes kurtosis differently; the function KURT(*data range*) computes "excess kurtosis" for sample data, which is CK − 3. (Excel does not have a corresponding function for a population).
- Thus, to interpret kurtosis values in Excel, distributions with values less than 0 are more flat, while those with values greater than 0 are more peaked.

 Slide - 38

Excel Descriptive Statistics Tool

This tool provides a summary of numerical statistical measures for sample data.

Data > Data Analysis > Descriptive Statistics

- Enter Input Range
- Labels (optional)
- Check Summary Statistics box

- The data must be in a single row or column. If the data are in multiple columns, the tool treats each row or column as a separate data set.

Example 4.28: Using the Descriptive Statistics Tool

- Purchase Orders database

Note: Results of the Analysis Toolpak do not change when changes are made to the data.

	A	B	C	D
1	*Cost per order*		*A/P Terms (Months)*	
2				
3	Mean	26295.31915	Mean	30.63829787
4	Standard Error	3078.053014	Standard Error	0.702294026
5	Median	15656.25	Median	30
6	Mode	14910	Mode	30
7	Standard Deviation	29842.8312	Standard Deviation	6.808993205
8	Sample Variance	890594573.8	Sample Variance	46.36238847
9	Kurtosis	2.079637302	Kurtosis	1.512188562
10	Skewness	1.664271519	Skewness	0.599265003
11	Range	127431.25	Range	30
12	Minimum	68.75	Minimum	15
13	Maximum	127500	Maximum	45
14	Sum	2471760	Sum	2880
15	Count	94	Count	94

Descriptive Statistics for Frequency Distributions

- Population mean: $\mu = \frac{\sum_{i=1}^{N} f_i x_i}{N}$ (4.16)
- Sample mean: $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{n}$ (4.17)
- Population variance: $\sigma^2 = \frac{\sum_{i=1}^{N} f_i (x_i - \mu)^2}{N}$ (4.18)
- Sample variance: $s^2 = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^2}{n - 1}$ (4.19)

Example 4.29: Computing Statistical Measures from Frequency Distributions

- Computer Repair Times

	A	B	C	D	E	F
1	Frequency Distribution Calculations					
2						
3	A/P Terms (x)	Frequency (f)	f*x	x - Mean	(x - Mean)^2	f*(x - Mean)^2
4	15	5	75	-15.6383	244.5563603	1222.781802
5	25	12	300	-5.6383	31.7904029	381.4848348
6	30	64	1920	-0.6383	0.407424174	26.07514713
7	45	13	585	14.3617	206.258488	2681.360344
8	Sum	94	2880			4311.702128
9						
10		Mean	30.6383		Sample variance	46.36238847

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

Grouped Data

- If the data are grouped into k cells in a frequency distribution, we can use modified versions of the formulas to estimate the mean and variance by replacing x_i with a representative value (such as the midpoint, M) for all the observations in each cell group and summing over all groups.

$$\mu = \frac{\sum_{i=1}^{k} f_i M_i}{N} \qquad (4.20)$$

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i M_i}{n} \qquad (4.21)$$

 Slide - 43

Example 4.30: Computing Descriptive Statistics for a Grouped Frequency Distribution

	A	B	C	D	E	F	G
1	**Grouped Frequency Distribution Calculations**						
2							
3	Cost/Order Group	Midpoint (x)	Frequency (f)	f*x	x - Mean	(x - Mean)^2	f*(x-Mean)^2
4	0 to 26000	13000	68	884000	-14936.17	223089180.6	15170064282
5	26000 to 52000	39000	8	312000	11063.8298	122408329.6	979266636.5
6	52000 to 78000	65000	11	715000	37063.8298	1373727478	15111002263
7	78000 to 104000	91000	4	364000	63063.8298	3977046627	15908186510
8	104000 to 130000	117000	3	351000	89063.8298	7932365776	23797097329
9		Sum	94	2626000			70965617021
10							
11			Mean	27936.2		Sample variance	763071150.8

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

 Slide - 44

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.

 Slide - 45

Example 4.31: Computing a Proportion

- Proportion of orders placed by Spacetime Technologies

$$= \frac{\text{COUNTIF(A4:A97, “Spacetime Technologies”)}}{94}$$

$$= \frac{12}{94} = 0.128$$

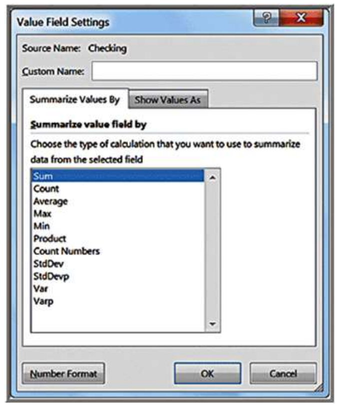
	A	B	C	D	E	F	G	H	I	J
1	**Purchase Orders**									
2										
3	**Supplier**	**Order No.**	**Item No.**	**Item Description**	**Item Cost**	**Quantity**	**Cost per order**	**A/P Terms (Months)**	**Order Date**	**Arrival Date**
4	Spacetime Technologies	A0111	6489	O-Ring	$ 3.00	900	$ 2,700.00	25	10/10/11	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

 Slide - 46

Statistics in PivotTables

Value Field Settings include several statistical measures:

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



Example 4.32: Statistical Measures in PivotTables

- *Credit Risk Data*
- First, create a PivotTable.
- In the *PivotTable Field List*, move Job to the *Row Labels* field and Checking and Savings to the *Values* field. Then change the field settings from “Sum of Checking” and “Sum of Savings” to the averages.

	A	B	C
1			
2			
3	Row Labels	Average of Checking	Average of Savings
4	Management	$606.94	$1,616.83
5	Skilled	$1,079.24	$1,836.43
6	Unemployed	$1,697.64	$2,760.91
7	Unskilled	$1,140.27	$1,741.44
8	Grand Total	$1,048.01	$1,812.56

Pearson Copyright © 2020, 2016, 2013 Pearson Education, Inc. All Rights Reserved

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.
- Sometimes, however, statistical relationships exist even though a change in one variable is not caused by a change in the other.

 Slide - 49

Measures of Association: Covariance

- **Covariance** is a measure of the linear association between two variables, *X* and *Y*. Like the variance, different formulas are used for populations and samples.
- Population covariance:

$$\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N} \quad \textbf{(4.25)}$$

  - Excel function: $=\mathrm{COVARIANCE.P}(\mathit{array1}, \mathit{array2})$
- Sample covariance:

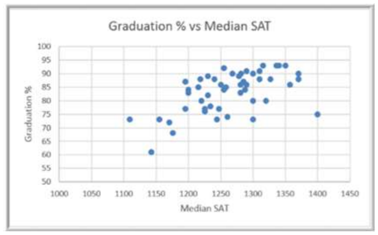
$$\operatorname{cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad \textbf{(4.26)}$$

  - Excel function: $=\mathrm{COVARIANCE.S}(\mathit{array1}, \mathit{array2})$
- The covariance between *X* and *Y* is the average of the product of the deviations of each pair of observations from their respective means.

 Slide - 50

Example 4.33: Computing the Covariance

- 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					Count	49
53					Covariance	263.3703231
54						
55					COVARIANCE.S	263.3703231

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:

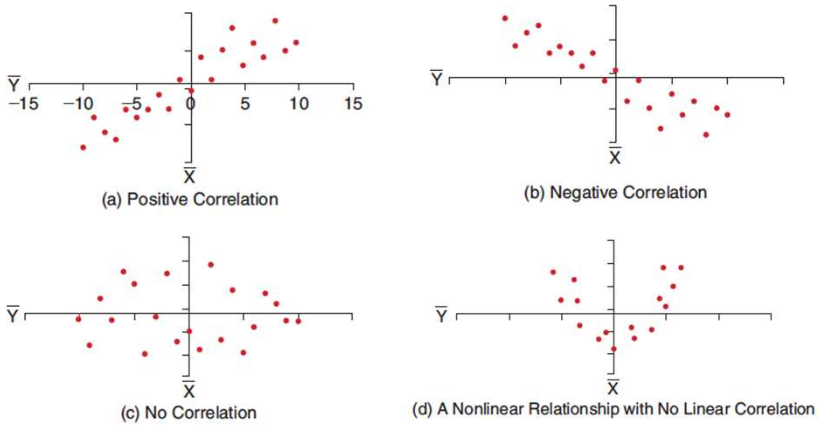
$$\rho_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (4.27)$$

- Correlation coefficient for a sample:

$$r_{XY} = \frac{\text{cov}(X, Y)}{s_X s_Y} \quad (4.28)$$

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

Examples of Correlation



Example 4.34: 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

Divide the covariance by the product of the standard deviations in cell F54.

Notes on the CORREL Function

- When using the CORREL function, it does not matter if the data represent samples or populations. In other words,

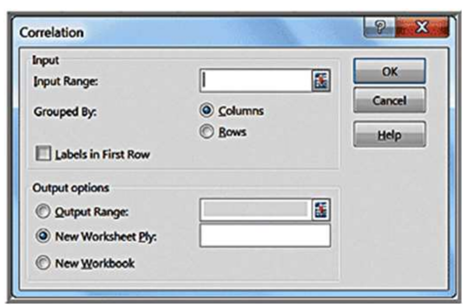
$$\text{CORREL}(array1, array2) = \frac{\text{COVARIANCE.P}(array1, array2)}{\text{STDEV.P}(array1) * \text{STDEV.P}(array2)}$$

and

$$\text{CORREL}(array1, array2) = \frac{\text{COVARIANCE.S}(array1, array2)}{\text{STDEV.S}(array1) * \text{STDEV.S}(array2)}$$

Excel Correlation Tool

Data > Data Analysis > Correlation



- Excel computes the correlation coefficient between all pairs of variables in the *Input Range*. *Input Range* data must be in contiguous columns.

Example 4.35: Using the Correlation Tool

- Colleges and Universities data

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

  - Moderate negative correlation between acceptance rate and graduation rate, indicating that schools with lower acceptance rates have higher graduation rates.
  - Acceptance rate is also negatively correlated with the median SAT and Top 10% HS, suggesting that schools with lower acceptance rates have higher student profiles.
  - The correlations with Expenditures/Student suggest that schools with higher student profiles spend more money per student.

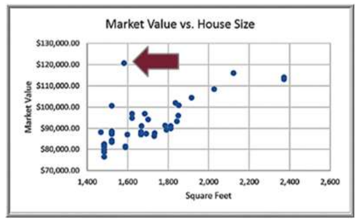
Identifying Outliers

- There is no standard definition of what constitutes an outlier.
- Some typical rules of thumb:
  - *z*-scores greater than +3 or less than −3
  - Values more than 3^{*}IQR to the left of Q_1 or right of Q_3 (extreme outliers)
  - Values between 1.5^{*}IQR and 3^{*}IQR to the left of Q_1 or right of Q_3 (mild outliers)

Example 4.36: 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.

 Slide - 59

Using Descriptive Statistics to Analyze Survey Data

- Descriptive statistics tools are extremely valuable for summarizing and analyzing survey data.
  - Frequency distributions and histograms for the ratio variables
  - Descriptive statistical measures for the ratio variables using the Descriptive Statistics tool
  - Proportions for various attributes of the categorical variables in the sample
  - PivotTables that break down the averages of ratio variables
  - Cross-tabulations
  - Z-scores for examination of potential outliers

 Slide - 60

Statistical Thinking in Business Decisions

- **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
- Work gets done in any organization through processes — systematic ways of doing things that achieve desired results.
- Understanding business processes provides the context for determining the effects of variation and the proper type of action to be taken.

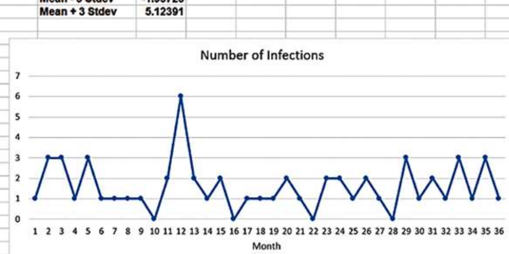
Example 4.37: Applying Statistical Thinking

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

Surgery Infections

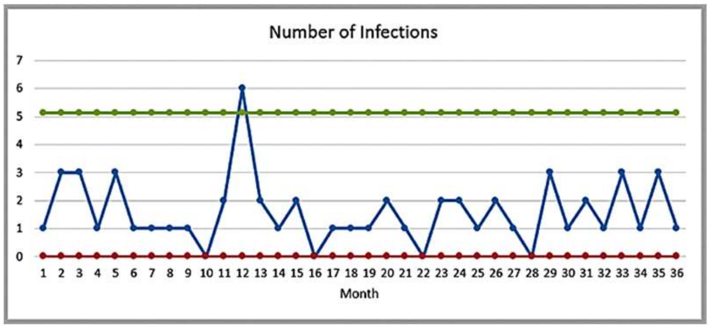
Month	Infections
1	1
2	3
3	3
4	1
5	3
6	1
7	1
8	1
9	1
10	0
11	2
12	6
13	2
14	1
15	2
16	0
17	1
18	1
19	1
20	2
21	1
22	0

Mean	1.58333
Standard Deviation	1.18019
Mean - 3 Stdev	-1.95725
Mean + 3 Stdev	5.12391



Example 4.37 Continued

- Three-standard deviation empirical rule:



- This suggests that month 12 is statistically different from the rest of the data.

 Slide - 63

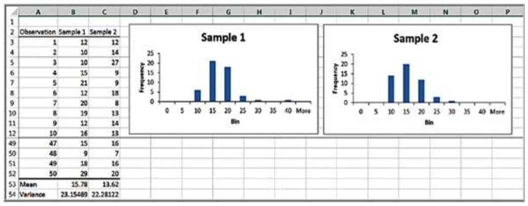
Variability in Samples

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

 Slide - 64

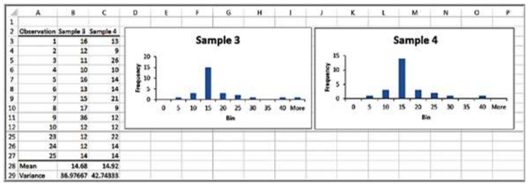
Example 4.38: Variation in Sample Data

- Samples from *Computer Repair Times* data
- Population statistics: $\mu = 14.91 \text{ days}, \sigma^2 = 35.5 \text{ days}^2$
- Two samples of size 50:



Observation	Sample 1	Sample 2
1	12	12
2	10	14
3	10	27
4	15	9
5	21	9
6	12	18
7	20	8
8	19	13
9	12	14
10	16	13
47	15	16
48	9	7
49	18	16
50	29	20
Mean	15.78	13.62
Variance	23.15469	22.28122

- Two samples of size 25:



Observation	Sample 3	Sample 4
1	16	13
2	12	9
3	11	26
4	10	10
5	16	14
6	13	14
7	15	21
8	17	9
9	36	12
10	12	12
23	12	22
24	12	14
25	14	14
Mean	14.68	14.92
Variance	36.97667	42.74333