Solutions Manual for Business Statistics Canadian 4th Edition by Black, Bayley, Castillo

ISBN: 9781119983187

ANSWERS TO CASES

Chapter 1
Canadian Farmers Dealing with Stress

In performing market research and other similar studies, it is important to properly identify target populations. In the case under study, the identification of the target population is important for the Canadian Agricultural Safety Association (CASA) because the farming industry represents a key sector of Western Canada's economy. Stress among farmers is a growing cause for concern and as such, accurate survey results will benefit the future of the industry. In this type of research, it is also important to identify the sampling frame, the type of survey to be conducted, the type of data to be collected, the level of measurement of the collected data, and any other pertinent statistic that will ensure that the results of the research that is to be conducted by Western Opinion Research Inc. on behalf of CASA are pertinent, reliable, and usable.

1. One population that was identified was the population of farmers across Canada. Western Opinion Research conducted the research and survey throughout Canada and used the population of Canadian farmers to obtain its results. There were no other populations that were contacted by the opinion firm. Instead of attempting to contact the entire population of Canadian farmers, the research group conducted their survey by using a sample from the population of interest. The survey was completed by 1100 farmers across Canada. The measurements obtained from the survey were generally qualitative and percentages were used to describe them. The results allowed the CASA to infer on potential consequences of the stress encountered by farmers, and used these results to initiate preventative actions that would at least stabilize the stress levels of farmers, but with the added intention of decreasing them. The inferences made with statistical results were imperative in offering stress counselling resources to farmers.

   The type of research that is conducted using the data obtained from studies such as the one commissioned by CASA allow organisations to use data analysis procedures in their normal course of business, whether it is for profit or not. The advantage of using inferential statistics, which are based on relevant samples, is that conclusions can be effectively drawn and which then pertain to the entire population under study, without having to conduct a census, which would most probably negatively impact the efficiency of the business operations.

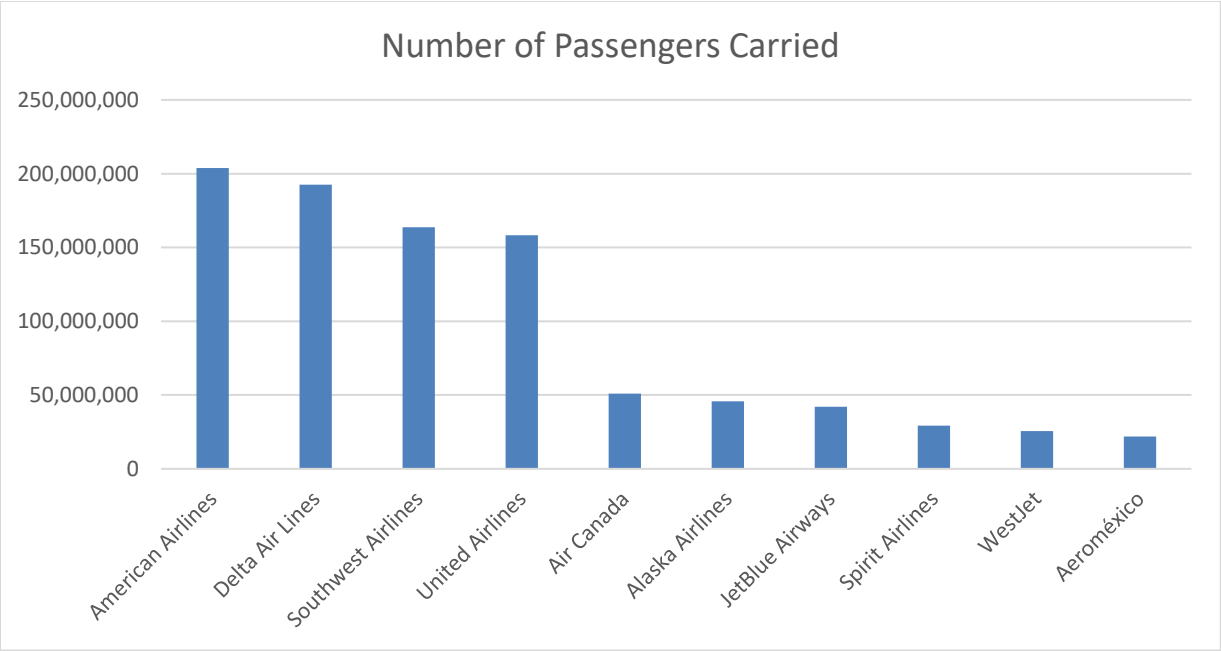
2.

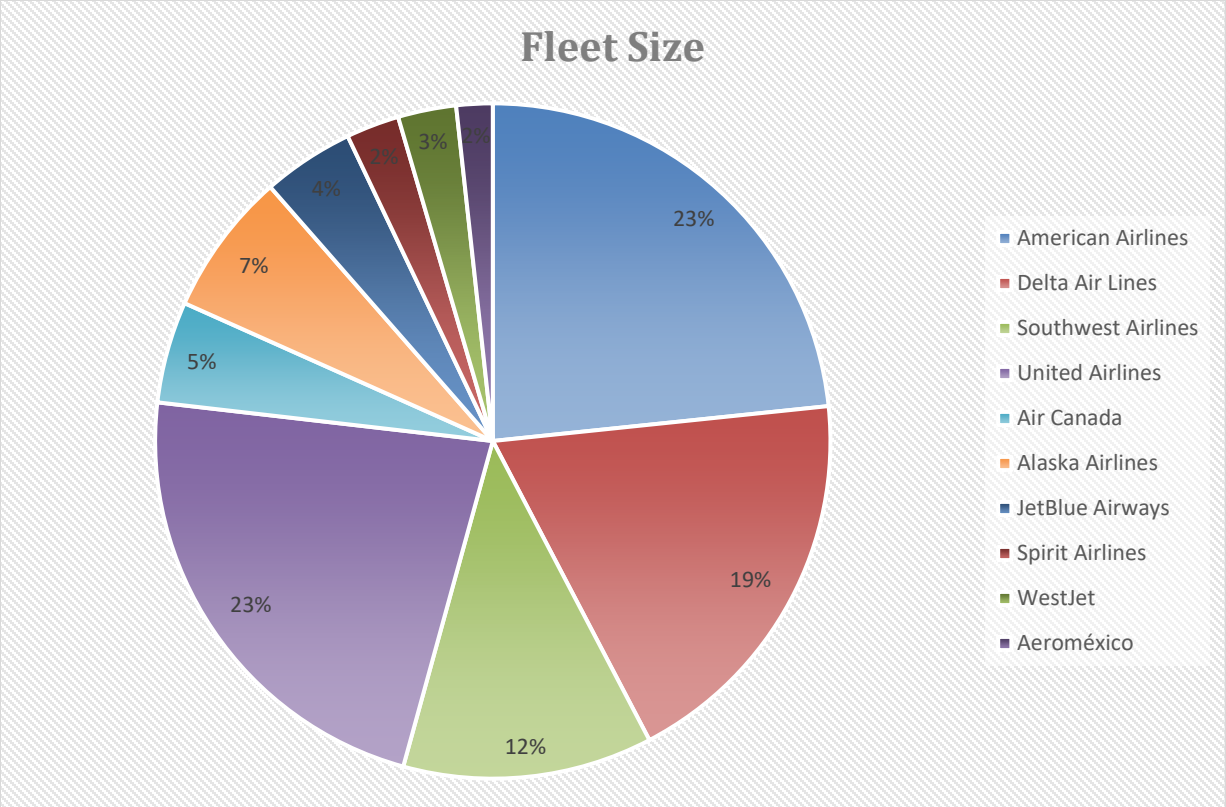
a. ranking of the level of stress	Ordinal level
b. number of farmers asking for help	Ratio level
c. number of farmers aware of help resources	Ratio level
d. number of farmers who try to manage stress	Ratio level
e. number of farmers interested in access to resources	Ratio level

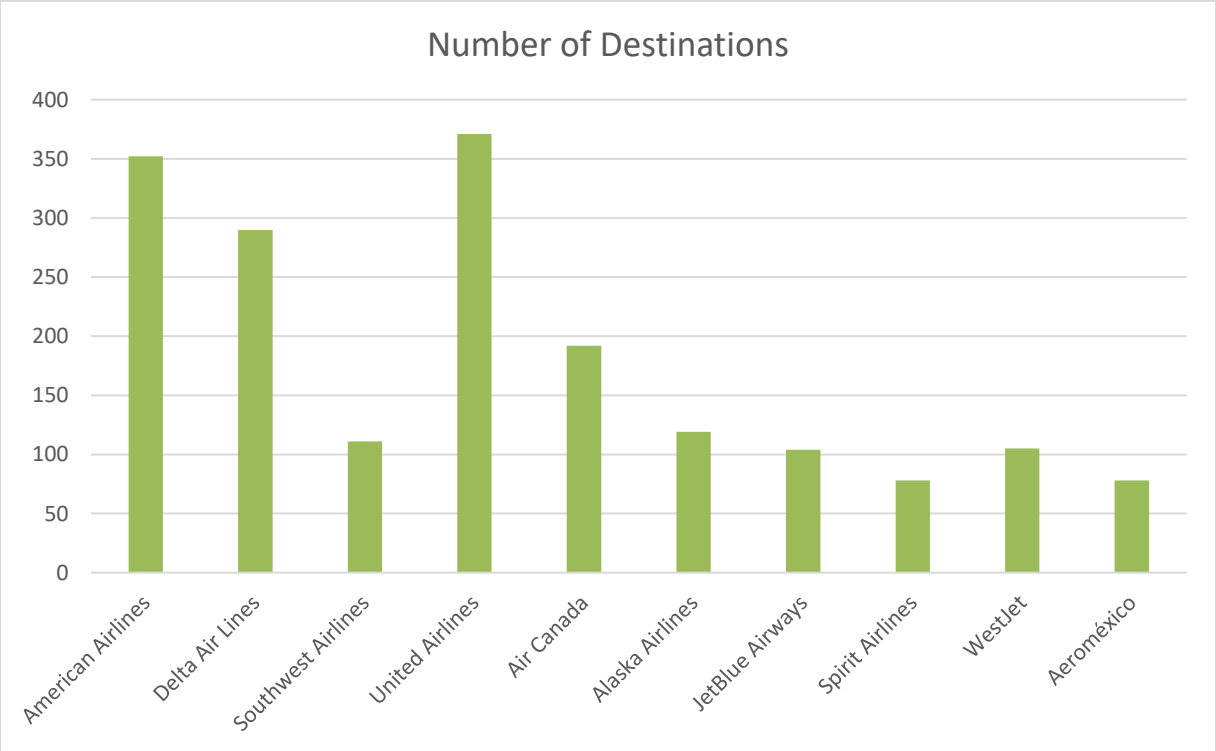
f. number of farmers nearly out of business	Ratio level
g. number of farmers who prefer dealing with stress on their own	Ratio level
h. number of farmers who prefer dealing with professionals on the phone	Ratio level
i. number of farmers who prefer dealing with professionals in person	Ratio level
j. age of respondent	Ratio level
k. gender of respondent	Nominal level
l. geographical region of respondent	Nominal level
m. time farmers spend dealing with stress	Ratio level
n. rating of stress-related factors	Ordinal level
o. rating of reasons for not seeking help for stress	Ordinal level

Chapter 2
Southwest Airlines and WestJet Airlines Ltd.

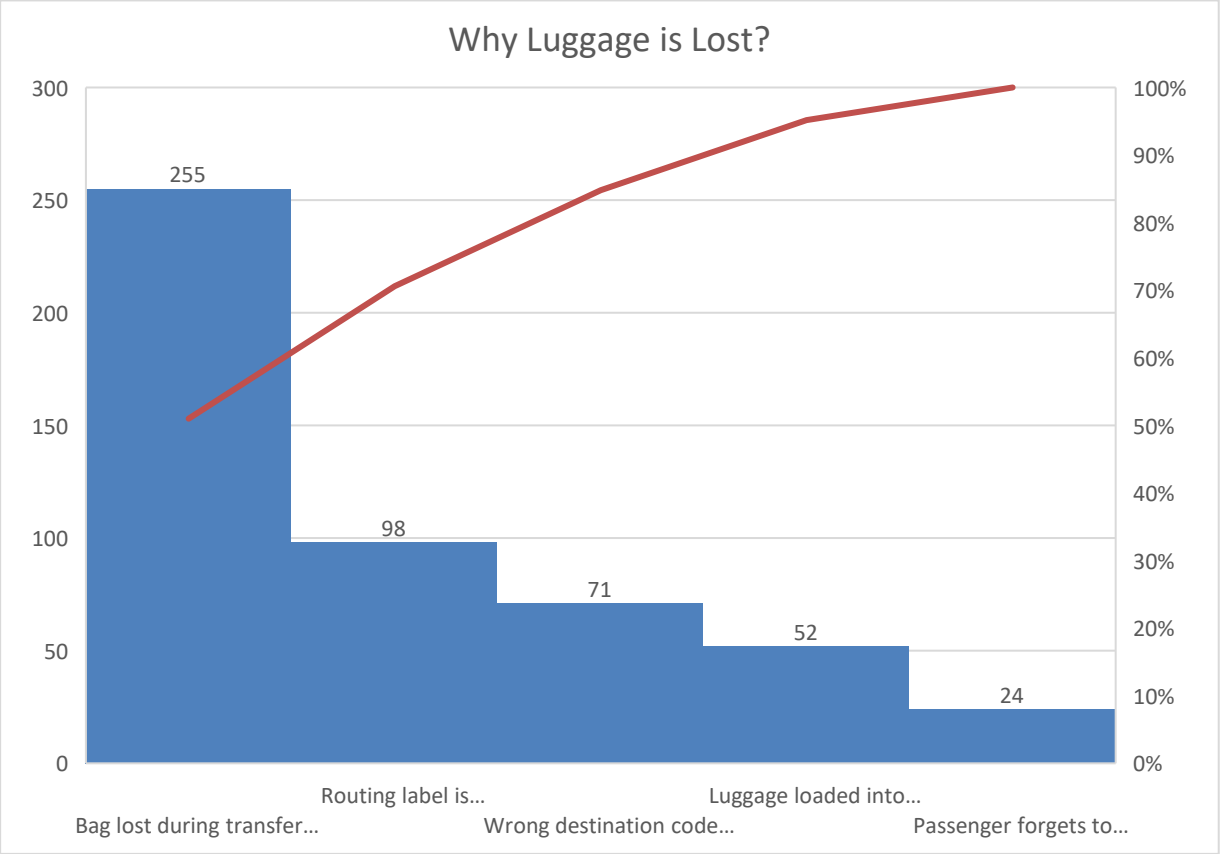
1. Shown below are three samples of visuals that could be generated to present the data. The student is encouraged to explore their own visualization options such as these and then write a brief report comparing WestJet to the other airlines.







2.
The Pareto Chart is a graphical technique for displaying problem causes. Question 2 of this case presents various causes for the problem of lost airline passenger luggage. Shown here is a Pareto Chart for the frequency of various causes associated with lost luggage.



The vertical bars of a Pareto Chart display the most common types of causes ranked in order of occurrence from left to right. This Pareto Chart shows that the number one cause of lost luggage is transfer from one plane to another (51% of lost luggage) followed by a damaged or lost routing label (an addition 20%). If an airline is undertaking an effort to significantly reduce the numbers of lost luggage, then they should start by tackling the "transfer" issue. In the quality improvement circles, this would be referred to as starting with the "low lying fruit." That is, if the transfer problem could be completely solved, an airline could reduce the numbers of lost luggage by 50%.

3.
Student answers will vary. There's a strong correlation between the number of passengers and the seats per aircraft, with two exceptions around 100 seats and 140 seats, both with extremely high passengers. There's a high, rather linear correlation between the number of passengers and the number of flights, with two exceptions around 600,000 flights and 750,000 flights, both with

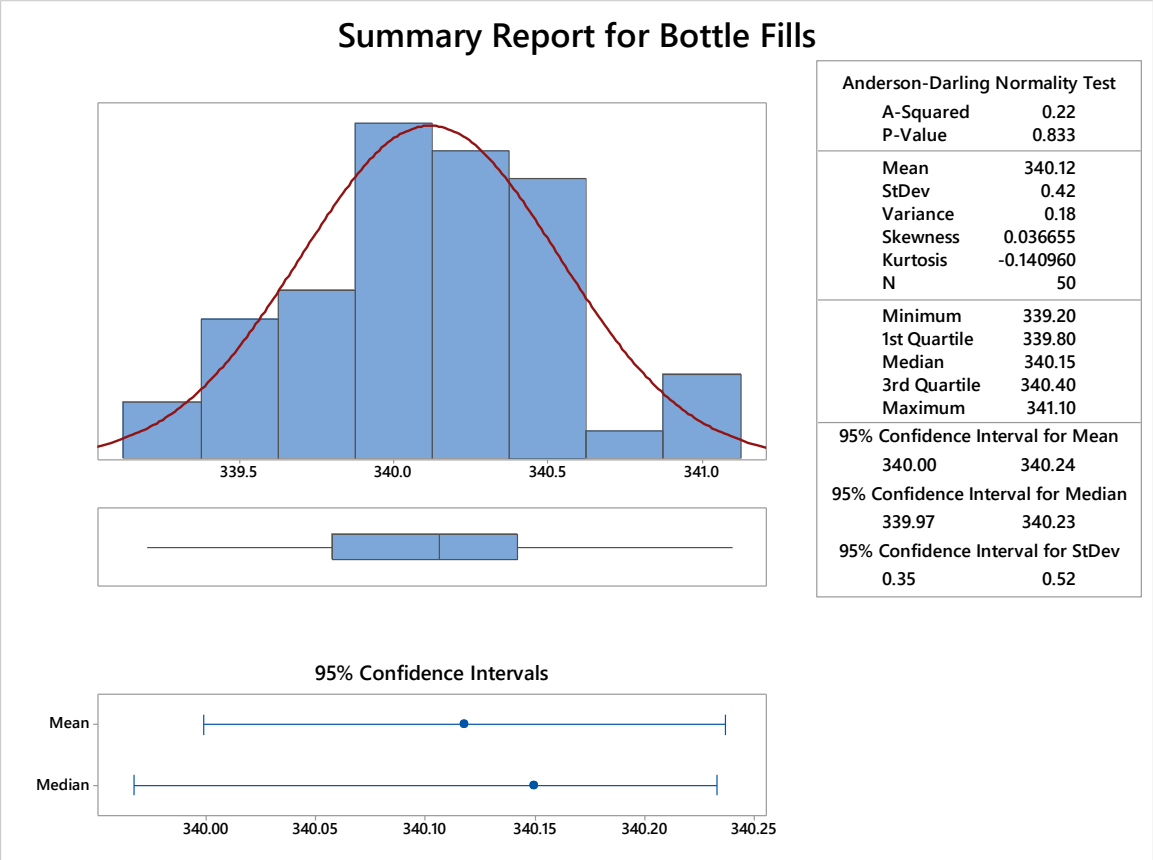
low passengers. For both of these apparent outliers, further investigation might be warranted to determine reasons for the extreme high and low number of passengers.

Chapter 3
Coca-Cola Develops the African Market

1. Shown below is output describing the sample of 50 bottle fills.

Descriptive Statistics: Bottle Fills

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
Bottle Fills	50	0	340.12	0.0592	0.419	339.20	339.80	340.15	340.40	341.10



Note that the mean fill is 340.12 oz. with a standard deviation of 0.42. The minimum fill is 339.20 oz. and the maximum fill is 341.10 oz. The median fill is 340.15 oz. The measure of skewness (.0367) demonstrates a very slight positive skewness. However, the

histogram and the high *p*-value associated with the normality test indicate that the data are approximately normally distributed. The mean fill of 340.12 oz. indicates that, on average, the sample fills are very near to 340 oz. and are, if anything, giving a slight amount of free product away to the consumer. Under the empirical rule, using $\mu = 340.12$ and $\sigma = .42$, 68% of the fills should be within $340.12 \pm .42$ or between 339.70 and 340.54 oz. and 95% of the fills should be within $340.12 \pm .84$ or between 339.28 and 340.96 oz.

2. The bottles have a label that claims there are 20 U.S. oz. of fluid therein. This sample of 150 bottles has an average of 20.008 oz. with a median of 19.997. The standard deviation of fills is 0.101 oz. The normality statistics and histogram overlaid with the normal curve indicate that the data are approximately normally distributed. We can apply the empirical rule. Approximately 68% of the fills are within $20.008 \pm 1(.101)$, from 19.907 to 20.109 oz.; and 95% within $20.008 \pm 2(.101)$, from 19.806 to 20.210.

The measure of skewness (.08) indicates very little skewness. The box plot indicates that there is an outlier at the lower end (extreme under filled bottle). Production and quality management people can better interpret these extreme fills in light of company goals and specifications. Overall, the fills are averaging very close to 20 oz., the fills are approximately normally distributed, and the standard deviation of fills is about .101 of an ounce.

Chapter 4
Bluewater Recycling Association Offers Bigger Bins

1. Here we have two events that have to be defined;

Let A = event that a St. Mary's resident is in the 45-64 age range;
Let B = event that a St. Marys resident (household) uses the new recycling bins.

The first probability is the probability that a St. Marys resident is in the 45-64 age range:

$P(A) = 0.28$

The second probability is a conditional probability, namely, the probability that a St. Marys resident is in the 45-64 age range given that he/she uses the new recycling bins:

$P(A|B) = 0.27$

The probability of having a St. Marys resident use the new recycling bins is provided in the text:

$P(B) = 0.94$

If age were independent of the initial use of the new recycling bins, we would have the following relationship:

$P(A) = P(A|B)$

Because the above relationship is not satisfied, it is clear that the two events are not statistically independent.

2. The probability that a randomly selected resident of St. Marys is either in the 45-64 age range or used the new bin during the initial two-month period is an application of the General Law of Addition:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

The probability $P(A \cap B)$ is obtained from a calculation using probabilities listed in Part 1, namely:

$$P(A \cap B) = P(A|B) \times P(B) = (0.27)(0.94) = 0.254$$

Therefore,

$$P(A \text{ or } B) = 0.28 + 0.94 - 0.254 = 0.966$$

3. Let A = Awareness through advertising; N = Awareness not through advertising;
Let P_i = Prior Probability

Event	Prior	$P(R\|P_i)$	$P(R \cap P_i)$	Revised
A	0.46	0.62	0.2852	0.646
N	0.54	0.29	0.1566	0.354
			P(R) = 0.4418	

The advertising campaign of the new recycling system was effective because the revised probability of the awareness campaign increased.

Chapter 5
Whole Foods Market Grows Through Mergers and Acquisitions

1. $n = 25$, $p = .30$.

   The expected number is:

   $\mu = n \cdot p = 25(.30) = 7.5$

   The probability that twelve or more have a high level of concern about food safety:

   Prob($x \geq 12 \mid n = 25$ and $p = .30$) =

   $${}_{25}C_{12}(.30)^{12}(.70)^{13} + {}_{25}C_{13}(.30)^{13}(.70)^{12} + {}_{25}C_{14}(.30)^{14}(.70)^{11} + \ldots + {}_{25}C_{25}(.30)^{25}(.70)^{0} =$$

   $.0268 + .0115 + .0042 + .0013 + .0004 + .0001 + ... + .0000 = .0443$

   If 30% of consumers have a high level of concern about food safety ($p = .30$), the expected number of purchasers from a sample of 25 is 7.5. Twelve or more are considerably more than the expected number (7.5). How often would one get twelve or more out of twenty-five when only 7.5 is expected? – about 4.43% (.0443) of the time. Therefore, if twelve or more out of twenty-five actually have a high level of concern about food safety, there is some evidence that the level of concern about food safety might actually be greater than 30%.

2. Let $\lambda = 3.4$ customers per minute. Shown below are some of the values for the Poisson distribution with $\lambda = 3.4$:

x	Probability
0	.0334
1	.1135
2	.1929
3	.2186
4	.1858
5	.1264
6	.0716
7	.0348
8	.0148
9	.0059
10	.0019
11	.0006
12	.0002

Store managers want to staff enough checkout lines such that only 1% of the time can demand not be met. Examining the probabilities for $\lambda = 3.4$, the sum of the probabilities for $x \geq 9$ is .0059 + .0019 + .0006 + .0002 = .0086 or 0.86%. Thus, if store management staffs enough lines to be able to handle 8 or fewer customers, they will be able to meet demand over 99% of the time (100% - 0.86%). Less than one percent of the time will they be unable to meet demand because 9 or more customers should only occur 0.86% of the time.

The $\lambda = 3.4$ is for one minute. However, the question being posed here is: What is the probability that 12 or more customers will want to check out in a two-minute period? In order to work this problem, we must double lambda to $\lambda = 6.8$ for two minutes. From Table A.3 in the Appendix, the following probabilities are obtained for $\lambda = 6.8$:

x	Probability
12	.0227
13	.0119
14	.0058
15	.0026
16	.0011
17	.0004
18	.0002
19	.0001
	.0448

The probability that $x \geq 12$ when $\lambda = 6.8$, is .0448.

3. Age Study:

This is a hypergeometric problem with:

$N = 30$, $n = 10$, $A = 17$, and $x \leq 3$

Out of 30 total workers (N), 17 (A) are at least 40-years-old. If we randomly sample 10 (n) of these workers, what is the probably that three or fewer (x) will be at least 40-years-old?

The probability is computed as:

$[({}_{17}C_3)({}_{13}C_7)]/({}_{30}C_{10}) + [({}_{17}C_2)({}_{13}C_8)]/({}_{30}C_{10}) + [({}_{17}C_1)({}_{13}C_9)]/({}_{30}C_{10}) + [({}_{17}C_0)({}_{13}C_{10})]/({}_{30}C_{10}) =$

$.0388 + .0058 + .0004 + .0000095 =$ **.0450**

Visible Minority study:

This is a hypergeometric problem with:

$N = 30$, $n = 10$, $A = 9$, and $x = 7$

The probability is computed as:

$[({}_{9}C_7)({}_{21}C_3)]/({}_{30}C_{10}) =$ **.0016**

Chapter 6
Mercedes Goes after Younger Buyers

1. Data provided:
   Mercedes C300E: μ = \$56,900; σ = \$3,991
   BMW 330E: μ = \$54,900; σ = \$3,379

   a.) Prob($x \geq 55{,}400 \mid \mu = 56{,}900$ and $\sigma = 3991$): = 0.3758

$$z = \frac{x - \mu}{\sigma} = z = \frac{55{,}4000 - 56{,}900}{3{,}991} = -0.3758$$

Using the Standard Normal Distribution table provided by the textbook, the area for z = -0.3758 is .1462.

Prob($x \geq 55{,}400$) = .5000 + .1462 = .6462 = **64.62%**

Almost 65% of Mercedes dealers would be priced out of competition with this BMW model.

Prob ($x > 56{,}900 \mid \mu = 54{,}900$ and $\sigma = 3379$):

$$z = \frac{x - \mu}{\sigma} = z = \frac{56{,}900 - 54{,}900}{3{,}379} = 0.59$$

Using the Standard Normal Distribution table provided by the textbook, the area for z = 0.59 is .2224.

Prob($x \geq 56{,}900$) = .5000 - .2224 = .2776 = **27.76%**

Almost 28% of the BMW dealers are pricing the BMW 330E more than the average price of the Mercedes C300E. If BMW dealers are pricing the 330E at the same price or higher than the average price for the Mercedes C300E, it means the cars will be directly competing on features other than price.

Prob($x < 54{,}900 \mid \mu = 56{,}900$ and $\sigma = 3991$):

$$z = \frac{x - \mu}{\sigma} = z = \frac{54{,}900 - 56{,}900}{3{,}991} = -0.50$$

Using the Standard Normal Distribution table provided by the textbook, the area for z = -0.5 is .1915.

Prob($x < 54{,}900$) = .5000 - .1915 = .3085 = **30.85%**

Approximately 31% of Mercedes dealers are pricing the C300E less than the average price of the BMW 330E.

2. $a = 2.2$ $b = 7.0$ $x_1 = 3.1$ $x_2 = 5.9$

Prob. = (5.9 – 3.1) / (7.0 – 2.2) = **58%**

$a = 2.2$ $b = 7.0$ $x_1 = 2.2$ $x_2 = 2.8$

Prob. = (2.8 – 2.2) / (7.0 – 2.2) = **12.5%**

$a = 2.2$ $b = 7.0$ $x_1 = 5.4$ $x_2 = 7.0$

Prob. = (7.0 – 5.4) / (7.0 – 2.2) = **33%**

3. $\lambda = 1.37$ cars/3 hours, $\mu = 1/1.37 = .73$ of 3 hours = 2.19 hours

For 1 hour:

1 hour = .333 of 3 hours. $x_0 = 0.333$. The cumulative probability of this time interval is .3663. This means that there is a 36.63% chance that there will be less than one hour between sales.
For 12 hours: 12 hours = 4 times 3 hours. $x_0 = 4$. The cumulative probability for this time interval is .9958 meaning that there is a 99.58% chance that there will be less than 12 hours between sales. The complement of this is that there is a 1 - .9958 = .0042 = 0.42% chance that there will be more than 12 hours between sales.

Managers know that there is an almost 37% chance of a sale within every hour. They need to determine how much staffing it takes to sell a car every hour or less. Given that it

takes several potential buyers and often multiple visits to the dealership to sell one car and that it is relatively likely (probability almost 75%) that they will close a sale every 3 hours ($x_0 = 1$), the dealership should never go without having salespeople around and may have to have several employees around all the time.

By having good exponential and Poisson distribution data, one can, to some extent, track the impact of advertising on sales by testing values of λ using random arrival data in time periods following advertising to determine if λ has increased. For example, if $\lambda = 1.37$ every 3 hours but following a advertising campaign, there is a randomly selected 3 hours period and 5 cars are sold, then management might be able to statistically justify that the λ has increased and then conclude that the advertising campaign was the cause.

In many businesses, the value of lambda changes with time of day, day of the week, and season of the year. In the car business, there may be an increase in sales on the weekend, in the evening, or perhaps in the fall when new models arrive. Students should always be cautioned about using the same value of lambda for all time periods. Many students know intuitively that lambda varies over time.

Chapter 7
3M

1. The answers to this question will vary. Students should select one of the four types of random sampling or a hybrid (e.g. area – stratified) to use in their sampling plan. The target population is that group of people to whom the researcher wants to infer. For example, if 3M wanted to determine the value of the index, usage rates, and general attitudes toward the index amongst all adults in the U.S., then all adults in the U.S. should be their target population. The frame is the list or roster of this target population from which the researchers sample. What is the frame of all adults in the U.S.? This is a difficult question. Many national lists pertain to some specialty group such as registered Republicans, Visa card users, Internet users, Catholic church members, etc. which really do not capture all U.S. adults. Instead of searching for a national frame, the researcher might want to use some form of area sampling such as selecting a test market city which is thought to be similar in demographics to the U.S. This might make frame identification easier because some test market city directories such as the phone book or voter registration list might be accessible and include most adults. On the other hand, it might make sense to use a list of all adopters of the Value Index Score as the frame and sampling from this frame, attempt to answer questions about the index's value and usage rates along with general attitudes towards the index. If such a frame were available, researchers might choose to stratify the population by user including providers, customers, patients, etc. along with geographic location, level of usage, length of time using the index, among others.

2. Prob.($\bar{x} > 16.0 \mid \mu = 15.3, \sigma = 4.5$, and $n = 35$):

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{16 - 15.3}{4.5/\sqrt{35}} = 0.92$$

From Table A.5, the area for $z = 0.92$ is .3212

Prob.($\bar{x} > 16.0$) = .5000 - .3212 = **.1788**

There is a 17.88% probability that the sample mean of more than 16.0 was obtained by chance. It is not conclusive from this that the population mean is longer 15.3. This is a good place for the instructor to mention .05 as a common standard for low probability and begin the groundwork for chapter 9.

Prob.($\bar{x} > 9.0$ | $\mu = 9.9$, $\sigma = 4.5$, and $n = 60$):

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{9 - 9.9}{4.5/\sqrt{60}} = -1.55$$

From Table A.5, the area for z = -1.55 is .4394

Prob.($\bar{x} > 9$) = .5000 + .4394 = **.9394**

There is a 93.94% probability that a sample mean of more than 9.0 was obtained by chance. This is a very likely event given that the population mean is 9.9.

Prob.($20.0 \leq \bar{x} \leq 22.0$ | $\mu = 20.3$, $\sigma = 4.5$, and $n = 43$):

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{22.0 - 20.3}{4.5/\sqrt{43}} = 2.48$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{20.0 - 20.3}{4.5/\sqrt{43}} = -0.44$$

From Table A.5, the area for z = 2.48 is .4934

From Table A.5, the area for z = -0.44 is .1700

Prob.($20.0 \leq \bar{x} \leq 20.8$) = .4934 + .1700 = **.6634**

There is a 66.34% probability that the sample mean of between 22.0 and 20.8 was obtained by chance. This is a very likely event given that the population mean is 20.3.

3. On a previous survey, 62% of device owners wanted better color, p = .62.

   A new survey of n = 450 is taken, what is the probability that more than 65% of these device owners want better color? That is, $\hat{p} > .65$?

   Prob.($\hat{p} > .65$ | $n = 450$ and $p = .62$):

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.65 - .62}{\sqrt{\frac{(.62)(.38)}{450}}} = 1.31$$

From Table A.5, the area for z = 1.31 is .4049.

Prob.($\hat{p} > .65$) = .5000 - .4049 = **.0951**

If 62% of device owners want better color, then the probability of sampling 450 such owners and having more than 65% want better color is .0951 or will occur about 9.51% of the time.

On a previous survey, 29% of device owners wanted more realistic colors, p = .29.

A new survey of n = 270 is taken, what is the probability that 25% or fewer of

these device owners want more realistic colors? That is, $\hat{p} \leq .25$?

Prob.($\hat{p} \leq .25 \mid n = 270$ and $p = .29$):

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.25 - .29}{\sqrt{\frac{(.29)(.71)}{270}}} = -1.45$$

From Table A.5, the area for z = -1.45 is .4265.

Prob.($\hat{p} \leq .25$) = .5000 - .4265 = **.0735**

If 29% of device owners want more realistic color, then the probability of sampling 270 such owners and having 25% or less want better color is .0735 or will occur about 7.35% of the time.

On a previous survey, 18% of device owners wanted bolder colors, p = .18.

A new survey of n = 950 is taken, what is the probability that between 152 and 200 want bolder colors?

152/950 = .16 and 200/950 = .21

What is the probability that between .16 and .21 of device owners want bolder colors?

Prob.($.16 \leq \hat{p} \leq .21 \mid n = 950$ and $p = .18$):

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.16 - .18}{\sqrt{\frac{(.18)(.82)}{950}}} = -1.60$$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{.21 - .18}{\sqrt{\frac{(.18)(.82)}{950}}} = 2.41$$

From Table A.5, the area for z = -1.60 is .4452.

From Table A.5, the area for z = 2.41 is .4920.

Prob.($.16 \leq \hat{p} \leq .21$) = .4452 + .4920 = **.9372**

If 18% of device owners want bolder colors, then the probability of sampling 950 such owners and having between 16% and 21% want bolder color is .9372. This is a very likely occurrence.

Chapter 8
Solutions Your Organized Living Store

1. $n = 115$ For 95% confidence, $z = 1.96$

Use: $\hat{p} \pm z\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$

1) Yes: $\hat{p} = \frac{73}{115} = .6348$

$$.6348 \pm 1.96\sqrt{\frac{(.6348)(.3652)}{115}} = .6348 \pm .0880$$

$.5468 \leq p \leq .7228$

2) Yes: $\hat{p} = \frac{81}{115} = .7043$

$$.7043 \pm 1.96\sqrt{\frac{(.7043)(.2957)}{115}} = .7043 \pm .0834$$

$.6209 \leq p \leq .7877$

3) Yes: $\hat{p} = \frac{88}{115} = .7652$

$$.7652 \pm 1.96\sqrt{\frac{(.7652)(.2348)}{115}} = .7652 \pm .0775$$

$.6877 \leq p \leq .8427$

4) Yes: $\hat{p} = \frac{66}{115} = .5739$

$$.5739 \pm 1.96\sqrt{\frac{(.5739)(.4261)}{115}} = .5739 \pm .0904$$

$.4835 \leq p \leq .6643$

2. $n = 21$ df = 20 For 95% confidence, $t_{.025,20} = 2.086$

Use: $\bar{x} \pm t\frac{s}{\sqrt{n}}$

1) $\bar{x} = 42.4$ s = 5.2

$$42.4 \pm 2.086\frac{(5.2)}{\sqrt{21}} = 42.4 \pm 2.37$$

$40.03 \leq \mu \leq 44.77$

2) $\bar{x} = 44.9$ $s = 3.1$

$$44.9 \pm 2.086\frac{(3.1)}{\sqrt{21}} = 44.9 \pm 1.41$$

$43.49 \leq \mu \leq 46.31$

3) $\bar{x} = 38.7$ $s = 7.5$

$$38.7 \pm 2.086\frac{7.5}{\sqrt{21}} = 38.7 \pm 3.41$$

$35.29 \leq \mu \leq 42.11$

4) $\bar{x} = 35.6 \qquad s = 9.2$

$$35.6 \pm 2.086 \frac{(9.2)}{\sqrt{21}} = 35.6 \pm 4.19$$

$31.41 \leq \mu \leq 39.79$

5) $\bar{x} = 34.5 \qquad s = 12.4$

$$34.5 \pm 2.086 \frac{12.4}{\sqrt{21}} = 34.5 \pm 5.64$$

$28.86 \leq \mu \leq 40.14$

6) $\bar{x} = 41.8 \qquad s = 6.3$

$$41.8 \pm 2.086 \frac{6.3}{\sqrt{21}} = 41.8 \pm 2.87$$

$38.93 \leq \mu \leq 44.67$

Chapter 9
A&W's New Menu Targets Meat Alternatives

1.

a)

$H_0 = 0.80$

$H_a \neq 0.80$

Let $\alpha = 0.05$ For a two-tailed test $\alpha / 2 = 0.025$

The critical value for this test is $z_{0.025} = \pm 1.96$

The sample size is: $n = 800$ $x = 615$

The sample proportion is: $\hat{p} = \frac{615}{800} = 0.7688$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} = \frac{0.7688 - 0.80}{\sqrt{\frac{(.80)(.20)}{800}}} = -2.21$$

Since the observed $z = -2.21 < z_{0.025} = -1.96$, the decision is to **reject the null hypothesis**. The proportion of Gen Z who eat meat alternatives is not 0.80. The sample data indicate that the proportion is lower than 0.80.

The probability of discrediting the claimed percentage (of 80%) if in fact it were true is **the probability of committing a Type I error**, also known as the level of significance, α. In this case, this probability is $\alpha = 0.05$.

b)

$H_0 = 0.50$

$H_a > 0.50$

The level of significance is not given in this case. We will use the p-value approach to test the above hypothesis.

The sample size is: $n = 500$ $x = 381$

The sample proportion is: $\hat{p} = \frac{375}{500} = 0.762$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} = \frac{0.7620 - 0.5}{\sqrt{\frac{(.5)(.5)}{500}}} = 11.717$$

The p-value is equal to the probability that the observed z-value exceeds 11.717. This probability is very close to 0. Hence the null hypothesis will be rejected for any level of significance, which is greater than the p-value, and as such, the test will be significant for practically any p-value. In conclusion, the sample result provides sufficient evidence to conclude that a higher proportion of consumers in the age range of 25 to 35 have increased their consumption of non-meat burger..

c)

$Ho: p = 0.16$
$Ha: p \neq 0.16$

Let $\alpha = 0.05$ For a two-tailed test $\alpha / 2 = 0.025$
The sample proportion of Canadian vegetarians who live in British Columbia obtained from the sample of 1,200 Gen Z folk is 0.1442. The 95% confidence interval is (0.1275, 0.1608); since the hypothesized value of 0.16 is within the confidence interval it is likely that the population proportion of Canadian vegetarians who live in British Columbia is 0.16. Using the hypothesis testing approach would result in the same conclusion as evidenced by the large p-value (0.1346). At a significance level of 0.05, since the p-value is greater than 0.05, the null hypothesis will not be rejected that the population proportion of Canadian vegetarians who live in British Columbia is still 0.16.

2. a) $H_0 = 32$
$H_a < 32$

Let $\alpha = 0.01$
With the given assumption that the distribution of the ean age of Millennial consumers who consume the Beyond Meat burger is normally distributed, and with the size of the sample at 28, it is appropriate to use the t-distribution to test the population mean, along with the fact that the population standard deviation is unknown.

For a left-tailed test, the observed $t_{0.01,27} = -2.473$. The value of the t-statistic, as given in the output provided is -1.713. Since, the observed $t = -1.713 < t_{0.01,29} = -2.473$, the decision is **not to reject the null hypothesis**. There is therefore insufficient evidence to conclude that the mean age of Millennial consumers who consume the Beyond Meat burger is less than 32 years old. The p-value for this test yields the same results, by definition. It can be seen from the computer output that the p-value is 0.0491, meaning that the test performed would be significant for any level of significance greater than 0.0491; in this case the level of significance is only 0.01.

b) $H_0 = 65$
$H_a < 65$

Let $\alpha = 0.05$

With the given assumption that the distribution of the number of Beyond Meat burgers Millennials purchase each year is a normally distributed variable, with the size of the sample at 30, and with the fact that the population standard deviation is unknown, it is appropriate to use the t-distribution to test the population mean.

For a left-tailed test, the observed $t_{0.05,29} = -1.699$.

The value of the test statistics is $\bar{x} = \frac{\sum x_i}{n} = \frac{1325}{30} = 44.17$

Because the population standard deviation is unknown, we will use the sample standard deviation as its estimate.

$$s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}} = \sqrt{\frac{68127 - \frac{(1325)^2}{30}}{29}} = 18.2002$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{44.17 - 65}{\frac{18.2002}{\sqrt{30}}} = -6.2686$$

Since the observed $t = -6.2686 < t_{0.05,19} = -1.699$, the decision is to **reject the null hypothesis**. The average number of frozen pizzas that Millenials purchase per year is much less than 65. The p-value for this test is less than 0.001, which corroborates the statistical decision of rejecting the null hypothesis.

Chapter 10
Seitz LLC: Producing Quality Gear-Driven and Linear-Motion Products

1. Comparing last year's mean transactions to this year's:

$n_1 = 20 \quad \bar{x}_1 = 2300 \quad s_1 = 500$
$n_2 = 25 \quad \bar{x}_2 = 2450 \quad s_2 = 540$

$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 \neq 0$

$df = n_1 + n_2 - 2 = 20 + 25 - 2 = 43 \quad \alpha = .05 \quad \alpha/2 = .025$

$t_{.025,43} = \pm 2.021$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} =$$

$$\frac{(2300 - 2450) - (0)}{\sqrt{\frac{(500)^2(19) + (540)^2(24)}{20 + 25 - 2}}\sqrt{\frac{1}{20} + \frac{1}{25}}} = \mathbf{-0.96}$$

Because $t = -0.96 > t_{.025,43} = -2.021$, the decision is to **fail to reject the null hypothesis**. There is not enough evidence here to say that there is any difference in the average dollar amount of sales between this year and last.

2. Comparison of tractors at two plants using a confidence interval:

$n_1 = 45 \quad x_1 = 18 \quad n_2 = 51 \quad x_2 = 12$

$$\hat{p}_1 = \frac{18}{45} = .4000 \qquad \hat{p}_2 = \frac{12}{51} = .2353$$

$$(\hat{p}_1 - \hat{p}_2) \pm z\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$(.40 - .2353) \pm 1.96\sqrt{\frac{(.40)(.60)}{45} + \frac{(.2353)(.7647)}{51}} = .1647 \pm .1845$$

$-.0198 \leq p_1 - p_2 \leq .3492$

The point estimate of the difference in quality of tractors at the two plants is 16.47%. However, due to the relatively small samples, the error of the interval is 18.45% which is greater than the point estimate. Combining the error of the interval with the point estimate results in the confidence interval shown above. Note that zero is in the interval indicating that there is a possibility that there is no difference in the quality ratings of tractors produced at the two plants. If this were a hypothesis testing problem, then the decision would be to fail to reject the null hypothesis based on the confidence interval's inclusion of zero.

3. 2020 vs. 2021:

$t = -1.85$ with a p-value of .066. This is not significant at $\alpha = .05$. There is no significant difference in the mean ratings between 2020 and 2021. This is underscored by the confidence interval that includes zero. However, if $\alpha = .10$ were used, there would be a significant difference. Examining the means reveals that the mean score for 2021 was higher. The sample sizes were 75 for 2013 and 93 for 2021.

4. Comparison of variances for week 1 and week 5:

H_0: $\sigma_1^2 = \sigma_2^2$
H_a: $\sigma_1^2 \neq \sigma_2^2$

Let $\alpha = .05$ $\alpha/2 = .025$ $df_1 = n_1 - 1 = 5$ $df_2 = n_2 - 1 = 6$

$F_{.025,5,6} = 5.99$ $F_{.975,6,5} = 1/5.99 = 0.167$

Week 1: $n_1 = 6$ $s_1 = 1.17$

Week 5: $n_2 = 7$ $s_2 = 1.68$

$$F = \frac{s_1^2}{s_2^2} = \frac{(1.17)^2}{(1.68)^2} = \mathbf{0.485}$$

Since the observed value of $F = 0.485$ is > the left tail critical value of $F = 0.167$, the decision is to fail to reject the null hypothesis. The variances of product being produced these two weeks are not significantly different. Management would probably like this because this indicative of consistent production patterns. Wide swings in variance would

be of concern because it would indicate that some weeks the variability is more out-of-control than others and a less consistent product is being produced.

Chapter 11
ASCO Valve Canada's RedHat Valve

1. The two by three factorial design is analyzed using a two-way ANOVA. There are two independent variables, temperature and supplier. Temperature has three treatment levels: 23°C, 49°C, and 68°C. Supplier has two classification levels: supplier 1 and supplier 2. The dependent variable is air pressure, measured in psi. Shown below is the Excel output for this analysis.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Supplier	88.88889	1	88.88889	0.006415	0.937484	4.747225
Temperature	98324.78	2	49162.39	3.547852	0.061585	3.885294
Interaction	1517.444	2	758.7222	0.054754	0.946953	3.885294
Within	166283.3	12	13856.94			
Total	266214.4	17				

First we examine the observed F for interaction which is 0.05 with a p-value of 0.9469. Since interaction is not significant for any commonly used significance level (α), we proceed to examine main effects. There is no significant difference between the two suppliers ($F = 0.006, p - value = 0.9375$). There is a significant in the strength of the new valve by temperature for any significance level greater than 0.0616. The mean psi for 23°C is 2355.167, for 49°C is 2348.167, and for 68°C is 2195. It appears that at 68°C, the valves are not as strong. Before concluding, it would appear appropriate to validate this conclusion by performing pairwise t-tests between each pair of temperature.

The results from Excel for each test are as follows:

A) t-Test: Two-Sample Assuming Equal Variances

	23C	*49C*
Mean	2355.167	2348.167
Variance	14232.17	15026.17
Observations	6	6
Pooled Variance	14629.17	
Hypothesized Mean Difference	0	
df	10	
t Stat	0.100242	
P(T<=t) one-tail	0.461067	
t Critical one-tail	1.812461	
P(T<=t) two-tail	0.922134	

t Critical two-tail	2.228139

B) t-Test: Two-Sample Assuming Equal Variances

	23C	*68C*
Mean	2355.167	2195
Variance	14232.17	4319.6
Observations	6	6
Pooled Variance	9275.883	
Hypothesized Mean Difference	0	
df	10	
t Stat	2.880415	
P(T<=t) one-tail	0.008187	
t Critical one-tail	1.812461	
P(T<=t) two-tail	0.016374	
t Critical two-tail	2.228139	

C) t-Test: Two-Sample Assuming Equal Variances

	49C	*68C*
Mean	2348.167	2195
Variance	15026.17	4319.6
Observations	6	6
Pooled Variance	9672.883	
Hypothesized Mean Difference	0	
df	10	
t Stat	2.69741	
P(T<=t) one-tail	0.011206	
t Critical one-tail	1.812461	
P(T<=t) two-tail	0.022413	
t Critical two-tail	2.228139	

The two-tail p-value for the test between the temperatures 23C and 49C is 0.9221, while the p-values for the other two tests are 0.0164 (for 23C vs 68C) and 0.0224 (for 39C vs 68C). Hence, the mean strengths are significantly different the latter two comparisons, and not for the first one (23C vs 49C). This clearly indicates that the mean strength is clearly different (lower) at 68°C. The implications of these results would indicate that under conditions of extreme heat, the valves might not be as strong as compared to temperatures in the vicinity of 50°C and below.

2. The data are analyzed using a one-way ANOVA. The independent variable is country with 4 classifications: Canada, Spain, Japan, and the U.S. The dependent variable is the cost reduction

incurred in using the energy-efficient Red hat valves. Shown below is the Excel output for this analysis.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	10.8375	3	3.6125	1.637394	0.22033	3.238872
Within Groups	35.3	16	2.20625			
Total	46.1375	19				

The results show that there is no significant difference in the cost reduction incurred among the four countries studied ($F = 1.64, p - value = 0.2203$). The management of ASCO should be confident that it can market the new Red hat valves in the four countries that were part of the analysis without having to implement special procedures in any one of the four countries, because, there is no statistical evidence that would suggest that any of the four countries would incur cost reductions which would be significantly different (lower) in any one of the four countries.

3. This test uses a randomized block design. The main independent variable of interest is the type of valve (two-way, three-way, or four-way). The blocking variable is the day of the week. Shown below is the Excel output for this analysis.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Day of the Week	0.609333	4	0.152333	1.432602	0.307639	3.837853
Type of valve	0.729333	2	0.364667	3.429467	0.084025	4.45897
Error	0.850667	8	0.106333			
Total	2.189333	14				

The results of the analysis indicate that lead time for the type of valve is not significantly different from one type to another, and this is valid for any level of significance greater than 0.0841. We can observe from the ANOVA table that the observed F-value is less than the F critical at $\alpha = 0.05$. Hence, the low confidence on any significant difference between type of valve. From the results of the analysis, each type of valve has a mean lead time of 1.62, 1.9, and 2.16 weeks. However, these means are not significantly different on the basis of the sample results.

To substantiate the above conclusion, the 95% confidence intervals for each of the three population mean lead times for all two-way, three-way, and four-way Red had valves are as follows:

for two-way valves: (1.159; 2.081)
for three-way valves: 1.589; 2.211)
for four-way valves: (1.835; 2.485)

All three confidence intervals overlap, which corroborates the conclusion that there is no significant difference between the mean lead time for each type of Red Hat valve.

Chapter 12
Caterpillar, Inc.

1. Shown next is Excel output from a regression analysis to predict the haul cost of a 12-cubic-yard end-dump vehicle by the speed of the vehicle.

SUMMARY OUTPUT For 12-Cubic-Yard Vehicle Model

Regression Statistics	
Multiple R	0.891
R Square	0.794
Adjusted R Square	0.753
Standard Error	0.341
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2.2457	2.2457	19.32	0.0070
Residual	5	0.5810	0.1162		
Total	6	2.8267			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2.3549	0.2973	7.921	0.0005
Speed	-0.0434	0.0099	-4.396	0.0070

The r^2 of the model, .794, is relatively high indicating strong predictability of haul cost by the speed of the vehicle. The *p*-value associated with both the overall model *F* value and the *t* test of the slope of the regression model is .007 indicating statistical significance at $\alpha = .01$ further supporting the strength of the model. The standard error of the regression model is 0.341 or 34.1 cents. Examining the residual output shown next, we can see that 6/7 = 85.7% of the residuals (errors) are less than .341.

Observation	*Residuals*
1	0.539
2	-0.064
3	-0.247
4	-0.290
5	-0.233
6	0.001
7	0.295

The regression model for the 12-cubic-yard end-dump vehicle is:

Haul Cost = 2.3549 -0.0434 Speed

The slope of the model indicates that there is a negative correlation between speed and haul cost. That is, higher speed indicates lower haul cost per cubic yard. As was mentioned in the case discussion question, faster speeds occur on flat, straight, and wide roads. In addition, faster highway speeds would result in shorter driving times, perhaps reducing the cost of labor, etc.

The haul cost for the 12-cubic-yard end-dump vehicle for 35 mph is:

Haul Cost = 2.3549 - 0.0434 Speed = 2.3549 – 0.0434 (35) = **0.8359**

The haul cost for the 12-cubic-yard end-dump vehicle for 45 mph is:

Haul Cost = 2.3549 - 0.0434 Speed = 2.3549 – 0.0434 (45) = **0.4019**

Shown next is Excel output from a regression analysis to predict the haul cost of a 20-cubic-yard bottom-dump vehicle by the speed of the vehicle.

SUMMARY OUTPUT For 20-Cubic-Yard Vehicle Model

Regression Statistics	
Multiple R	0.884
R Square	0.781
Adjusted R Square	0.737
Standard Error	0.285
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.4408	1.4408	17.80	0.0083
Residual	5	0.4047	0.0809		
Total	6	1.8455			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	1.8805	0.2481	7.579	0.0006
Speed	-0.0348	0.0082	-4.219	0.0083

The r^2 of the model, .781, is relatively high indicating strong predictability of haul cost by the speed of the vehicle. The *p*-value associated with both the overall model *F* value and the *t* test of the slope of the regression model is .008 indicating statistical significance at $\alpha = .01$ further supporting the strength of the model. The standard error of the regression model is 0.285 or 28.5 cents. Examining the residual output shown next, we can see that 6/7 = 85.7% of the residuals (errors) are less than .285.

Observation	*Residuals*
1	0.447
2	-0.049
3	-0.205
4	-0.242
5	-0.188
6	-0.020
7	0.257

The regression model for the 20-cubic-yard end-dump vehicle is:

$$\text{Haul Cost} = 1.8805 - 0.0348 \text{ Speed}$$

The slope of the model indicates that there is a negative correlation between speed and haul cost. That is, higher speed indicates lower haul cost per cubic yard. As was mentioned in the case discussion question, faster speeds occur on flat, straight, and wide roads. In addition, faster highway speeds would result in shorter driving times, perhaps reducing the cost of labor, etc.

The haul cost for the 20-cubic-yard bottom-dump vehicle for 35 mph is:

$$\text{Haul Cost} = 1.8805 - 0.0348 \text{ Speed} = 1.8805 - 0.0348\,(35) = \mathbf{0.6625}$$

The haul cost for the 20-cubic-yard bottom-dump vehicle for 45 mph is:

$$\text{Haul Cost} = 1.8805 - 0.0348 \text{ Speed} = 1.8805 - 0.0348\,(45) = \mathbf{0.3145}$$

2. Shown next is Excel output for a regression model to predict Caterpillar's Sales and Revenue by Year (see Section 12.9).

SUMMARY
OUTPUT

Regression Statistics	
Multiple R	0.060
R Square	0.004
Adjusted R Square	-0.073
Standard Error	9.04327
Observations	15

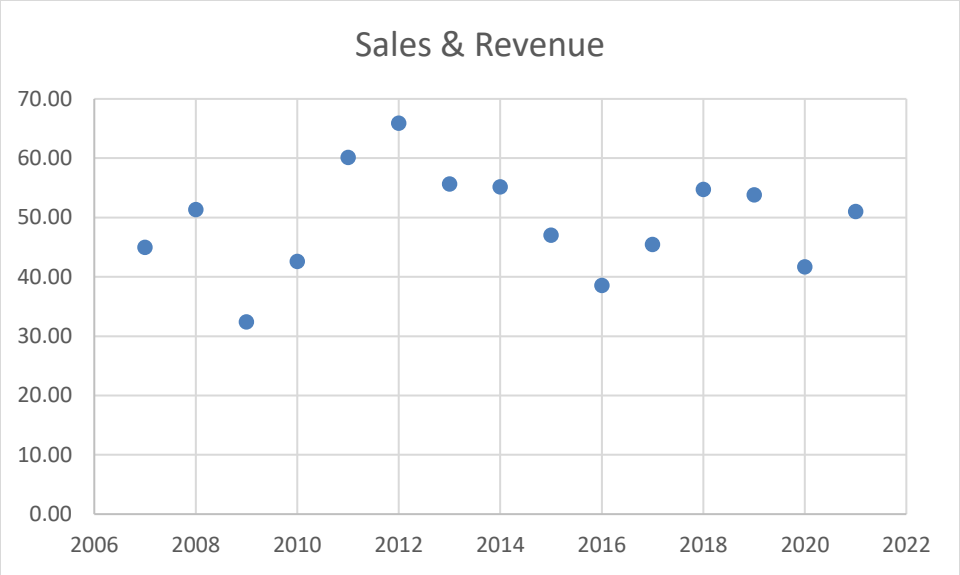
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.821	3.821	0.047	0.832
Residual	13	1063.149	81.781		
Total	14	1066.970			

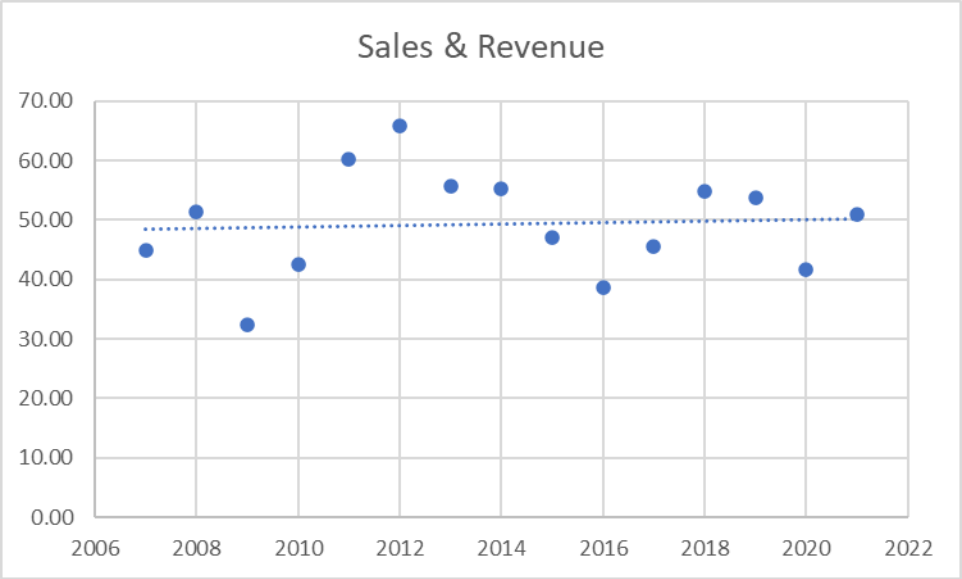
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-185.921	1088.446	-0.171	0.867
Year	0.117	0.540	0.216	0.832

The R^2 for the model is very low at .004 indicating that this is a very weak regression model.

The p-values of .832 for the overall F test and the t test for the slope are not significant for $\alpha = .05$. Examining the Excel-produced scatter plot shown below, it appears that the trend more curvilinear than linear.



Adding a trend line to the Excel scatter plot as shown below, highlights the notion that a line does not fit the graph well and that some form of curvilinear model might work better.



The regression model for the trend line is:

$$\text{Sales} = -185.921 + 0.117\ \text{Year}$$

An interpretation of this model is that, on average, sales and revenues are increasing \$0.117 billion per year. In this particular case, the *y*-intercept has no practical meaning other than if the company would have been in business in the year 0, it would have been losing a lot of money!

This model could be used to predict sales and revenues for Caterpillar in the year 2022:

Sales = --185.921 + 0.117 Year = --185.921 + 0.117(2024) =
$50.887 billion

The student can experiment in Excel with placing nonlinear models through the data in an attempt to explore for better fits. In chapter 14, the text presents polynomial fit models.

Chapter 13
Starbucks Introduces Debit Card

1. This model uses four independent variables in an effort to predict the amount of money people spend on their debit card. Overall, the model has modest to good predictability with an R^2 of .755 and a standard error of $22.15. This standard error indicates that about 95% of the time, the model will be within $\pm$2($22.15) or $\pm$$44.30 of the actual figure which is not particularly good. While the overall test of the model is significant (F = 15.38, p-value = .000007), an examination of the t tests and their associated p-values shows that only one of the predictors, income (t = 6.69, p-value = .000002) is significant. None of the other variables are even close. Had a simple regression model been developed using just income to predict the amount of the prepaid card, the R^2 would be .723, the t value for income would increase to 7.74, the standard error of the estimate would reduce to $21.96, and the overall F test would increase to 59.90.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.869
R Square	0.755
Adjusted R Square	0.706
Standard Error	22.1483
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	30175.0423	7543.7606	15.38	0.000007
Residual	20	9810.9577	490.5479		
Total	24	39986			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-83.826	22.494	-3.73	0.0013
Age	0.237	0.576	0.41	0.6852
Days	1.190	1.474	0.81	0.4291
Cups	1.422	2.631	0.54	0.5949
Income	2.407	0.360	6.69	0.000002

2. This model attempts to predict the number of days per month that a customer frequents Starbucks. The predictor variables are age, income, and number of cups of coffee per day. The Excel results of this analysis are shown below. The model is modest to weak with an R^2 of just .416. The standard error of 3.28 days indicates that the model would predict within $\pm$ 2(3.28) or $\pm$ 6.56 days about 95% of the time. A perusal of the data shows that the range of number of days is 16 days. The relatively large size of the standard error to this range is further evidence of the model's weakness. A study of the *t* statistics reveals that the predictor variable, cups, is the only significant predictor (t = 3.40, p-value .0027). The number of cups of coffee that a person drinks per day seems to be a good predictor of the number of times per month the person frequents Starbucks. Heavy coffee drinkers come often (the coefficient indicates a positive relationship between cups and frequency). In attempting to increase store traffic, Starbucks could target their marketing efforts at the more heavy coffee drinkers or develop and market products that might lure lighter coffee drinkers to their outlets for different reasons. If a simple regression model is used to predict number of days by cups of coffee, the R^2 is .345 and the standard error is 3.32.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.645
R Square	0.416
Adjusted R Square	0.332
Standard Error	3.2791
Observations	25

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	160.7593	53.5864	4.98	0.0091
Residual	21	225.8007	10.7524		
Total	24	386.56			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	5.9684	3.0651	1.95	0.0650
Cups	1.0644	0.3127	3.40	0.0027
Income	0.0716	0.0509	1.41	0.1742
Age	-0.0785	0.0835	-0.94	0.3578

3. Shown below is the output from an Excel multiple regression analysis to predict sales revenue by number of stores, number of drinks, and average weekly earnings. The predictability is extremely high with an R^2 of .9998. In predicting sales revenues that range from 400 to 2600, the standard error of the estimate is only 16.69. The overall F of 4539.21 is significant at alpha = .00001. While number of stores is not a significant predictor (t = -0.95, p-value = .41145), both number of drinks (t = -7.47, p-value = .00497) and average weekly earnings (t = 13.70, p-value .00084) are significant at α = .01. Notice that for the predictor, number of drinks, both the t value and the coefficient are negative. This indicates that, at least in this model with other variables in the model, there is a negative relationship between number of drinks and sales revenue. However, a cursory examination of the raw data shows that as sales revenues increase so do the number of drinks. This points out one of the dangers in over interpreting the regression coefficients (discussed in Chapter 14 in section on multicollinearity). When a simple regression model is run using number of drinks as the sole predictor of sales revenue, the r^2 is .929, and more importantly the regression coefficient is *positive* as is the t statistic. This might serve as an informal/intuitive introduction to the notion of collinearity. The correlation between the two significant predictors in the multiple regression model, number of drinks and weekly earnings, is .984.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9999
R Square	0.9998
Adjusted R Square	0.9996
Standard Error	16.689
Observations	7

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3792735.879	1264245.293	4539.21	0.000006
Residual	3	835.55	278.52		
Total	6	3793571.429			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-13500.237	946.164	-14.27	0.00075
Stores	-0.0264	0.0278	-0.95	0.41145
Drinks	-75.20448212	10.07005905	-7.47	0.00497
Earnings	38.989	2.846675638	13.70	0.00084

Chapter 14
Ceapro Turns Oats into Beneficial Products

1. Shown below is the Excel multiple regression output which contains a model to predict total size of purchase (expressed in thousands of dollars) using three predictors: Company size (in millions of sales), cost of delivery (in dollars), and number of similar products. Below that is a stepwise regression analysis for the same data. The full multiple regression model has an R^2 of 77.3%. However, the adjusted R^2 is only 71.1% which indicates there are some non significant predictors in the model. the analysis of the p-values shows that only company size makes a useful contribution to the overall model, while the number of similar products could be useful for any significance level greater than 7.5%. Cost of delivery is clearly non-significant (p-value = 0.9481). The overall F value is significant at $\alpha = 0.01$. The stepwise regression analysis confirms the usefulness of only one variable, namely, company size. The R^2 value of the stepwise regression model is 68.5%, all accounted for by only one variable. This result is consistent with the analysis made with the overall model. In conclusion, the most effective model in predicting size of purchase is a single regression model with company size as the only predictor: Size of purchase = 23.90411 + 1.78177 (Company Size), with size of purchase expressed in thousands of dollars and company size in millions. The positive sign for the regression coefficient confirms that the larger the size of the company, the larger the size of purchase; the cost of the delivery and the number of similar products do not seem to have any significant impact on the size of purchase for any significance level lower than 7.5%.

Overall Multiple Regression Model

Regression Statistics	
R-Square	0.77313
Adjusted R-Squared	0.71125
Standard Error	93.5374
Observations	15.00

ANOVA

	d.f.	SS	MS	F	p-value
Regression	3	327,966.21	109,322.07	12.49504	0.00073
Residual	11	96,241.63	8,749.24		
Total	14	424,207.84			

	Coefficients	Std Err	LCL-95%	UCL-95%	t Stat	p-value
Intercept	133.09852	65.51606	-11.10135	277.2984	2.03154	0.06707
Company Size ($ millions sales)	1.26347	0.54806	0.0572	2.46975	2.30535	0.04164
Cost of Delivery ($)	0.03024	0.06361	-0.10978	0.17025	0.47532	0.64386
Similar Products	-44.5949	21.73338	-92.42974	3.23995	-2.05191	0.06476

Table of Results for General Stepwise

X1 entered.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	290631.4453	290631.4453	28.28500404	0.000139381
Residual	13	133576.392	10275.10708		
Total	14	424207.8373			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	23.9041119	41.82544623	0.571520786	0.577394899	-66.454271	114.2624948
Company Size	1.781774712	0.335023035	5.31836479	0.000139381	1.05800145	2.505547974

No other variables could be entered into the model. Stepwise ends.

These findings may also beconfirmed when individual simple regression models are run for each independent variable, the summarized results of which are presented below:

Independent variables	Company Size ($ millions sales)
Regression Statistics	
R-Squared	0.68512
Adjusted R-Squared	0.66089

ANOVA

	d.f.	SS	MS	F	p-value
Regression	1	290,631.45	290,631.45	28.285	0.00014
Residual	13	133,576.39	10,275.11		
Total	14	424,207.84			

	Coefficients	Std Err	LCL 95%	UCL 95%	t Stat	p-value
Intercept	23.90411	41.82545	-66.45427	114.26249	0.57152	0.57739
Company Size ($ millions sales)	1.78177	0.33502	1.058	2.50555	5.31836	0.00014

Independent variables	Cost of Delivery ($)
Regression Statistics	
R-Squared	0.35247

Adjusted R-Squared	0.30266

ANOVA

	d.f.	SS	MS	F	p-value
Regression	1	149,521.78	149,521.78	7.07638	0.01963
Residual	13	274,686.05	21,129.70		
Total	14	424,207.84			

	Coefficients	Std Err	LCL 95%	UCL 95%	t Stat	p-value
Intercept	79.8599	57.97805	-45.39406	205.11385	1.37742	0.19164
Cost of Delivery ($)	0.16495	0.06201	0.03099	0.29891	2.66015	0.01963

Independent variables	Similar Products

Regression Statistics

R-Squared	0.4007
Adjusted R-Squared	0.3546

ANOVA

	d.f.	SS	MS	F	p-value
Regression	1	169,979.13	169,979.13	8.69189	0.01131
Residual	13	254,228.71	19,556.05		
Total	14	424,207.84			

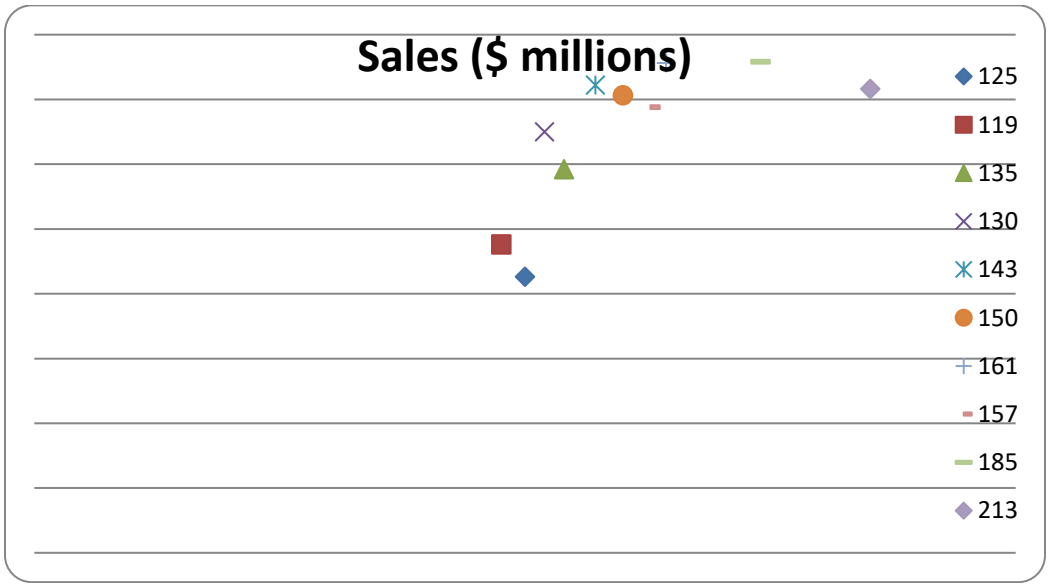
	Coefficients	Std Err	LCL 95%	UCL 95%	t Stat	p-value
Intercept	341.84697	60.85884	210.36944	473.3245	5.61705	0.00008
Similar Products	-80.24091	27.2169	-139.03945	-21.44237	-2.9482	0.01131

The independent variable "Company Size" has the strongest R^2 value, and is significantly larger than the R^2 values for the remaining two independent variables. As a next step, and without having stepwise regression available, would be to run a multiple regression by adding the independent variables in order of the R^2 beginning with the greatest value and working towards the lowest value. Once the Adjusted R^2 for the multiple regression model is no longer increasing then adding variables should stop. This is confirmed through the use of the stepwise regression model.

2. The regression analysis attempts to develop a model that can be used to predict average sales (in $ million) on the basis of two predictors, hours worked per week and number of customers. The scatter plots of each of these two predictors have a slight linear shape, but each display a somewhat similar shape to the upper left quadrant of Turkey's 4-quadrant approach. Neither simple regression models using hours worked per week or number of customers by themselves produced a significant t-statistic, and the r^2 values of each were quite low, indicating that neither variable is significant by itself as a predictor. The simple regression model using hours worked per week has a r^2 value of 6.7% and the p-value of the regression coefficient is 0.4707. The simple regression model using number of customers has a r^2 value of 5% and the p-value of the regression

coefficient is 0.5337.A multiple regression analysis was conducted using both variables as predictors, and the resulting model has a R^2 value of 17.05%, with both variables being statistically not significant (the p-values are 0.3473 and 0.3808 for hours worked per week and number of customers, respectively). An analysis of the correlation between the two predictor gives a correlation coefficient of -0.315, which does not indicate much collinearity between the predictors. The overall analysis seems to suggest that the one possibility for this model to work is by adding more data (observations), perhaps using quarterly data instead of annual data, or possibly, monthly data, if it is available. Also, the use of a non-linear regression model should be considered and explored to see if there would be a better fit than the linear model currently shows.

3. The scatter plot of the Sales and Number of Employees is shown below.



From the above graph (with sales on the vertical axis and number of employees on the horizontal axis, notice how the graph rises and then flattens out. This pattern fits closely with the upper left quadrant of Tukey's 4-quadrant approach. From this, and before attempting any model transformation, we performed a simple regression analysis using number of employees as predictor. The model obtained is:

$y = 11.7064 + 0.13665x$ where the p-value of the regression coefficient is 0.0288, indicating that number of employees makes a significant contribution to the prediction of sales; however, this simple regression model has a r^2 value of only 46.9%. We attempted to transform the model by adding a second predictor, the log of the number of employees, which is consistent with the patter of the scatter plot and as suggested by Tukey's 4-quadrant approach. The new model obtained is:

$y = -803.9 - 1{,}1097x_1 + 462x_2$ where x_1 represents the number of employees and x_2 represents the log of the number of employees. The p-values for both predictors are 0.0136 and 0.0077, respectively. it is also interesting to note the the inclusion of a second predictor has significantly increased the R^2 value to 82.02%, and as such, the predictability of the transformed model has also increased significantly. However, management must be careful in using this model and other

analyses (such as optimization) must be performed in order to obtain a solid understanding of the range of employees it can hire before the model reaches an optimum value. It can be easily verified that the model just obtained will peak in expected sales at $39.2 million when the number of employees is 180. Hiring beyond the 180 emplyeee figure will generate expected sales lesss than $39.2 million. furthermore, it can be verified that hiring to a level of 320 employees will not generate any expected sales, based on this model and its parameters. This means that the range of the x-value has to be less than 320.

Chapter 15
Dofasco Changes Its Style

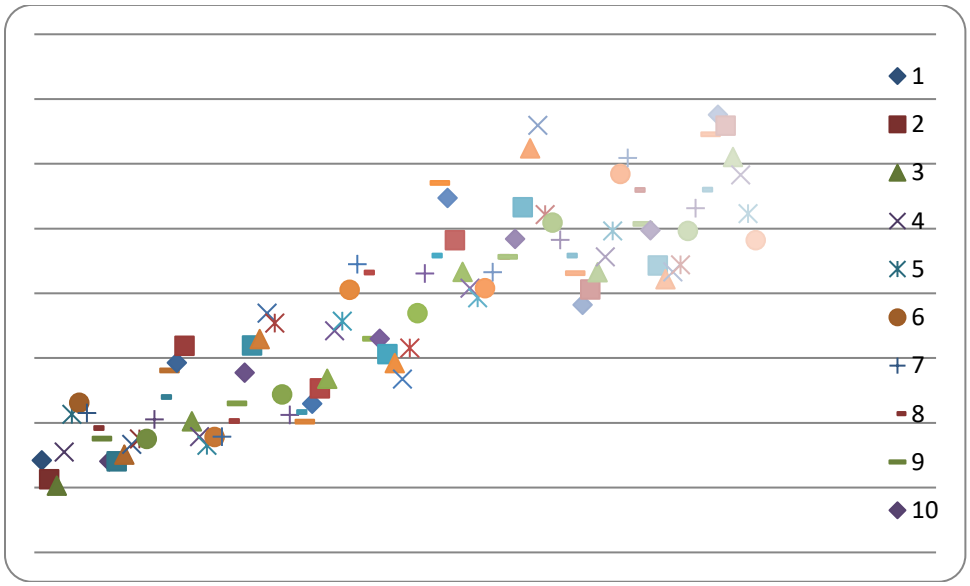
1. The decomposition analysis shows several elements of the Dofasco monthly sales data. First, the seasonal indices for each of the 12 months are as follows:

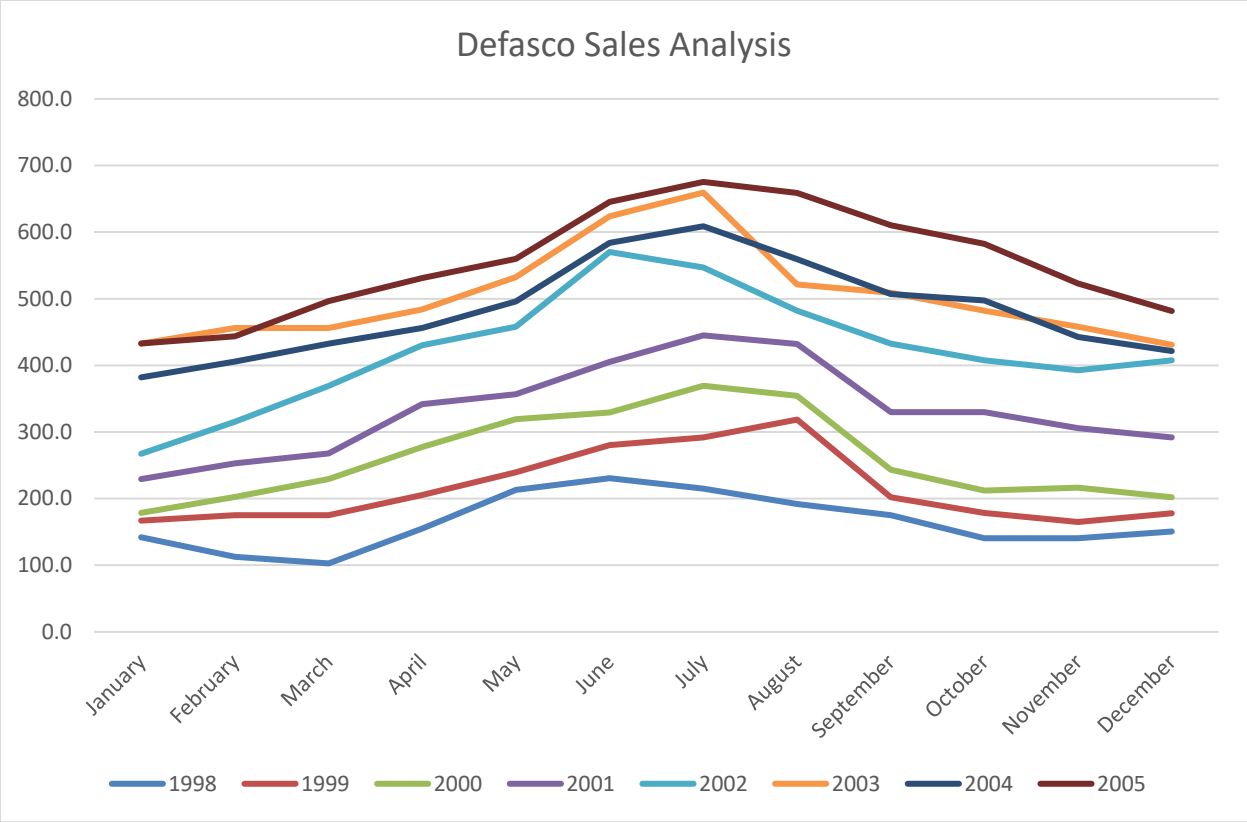
January	February	March	April	May	June
80.7978	85.9692	90.9959	102.4842	110.7161	126.0836

July	August	Sept	Oct	Nov	Dec
132.4183	121.698	97.3592	88.0568	82.5908	80.8301

There appears to be seasonal effects in the data. For example, the seasonal indices are less than 100 from january to march and from September to December. Dofasco sales peak from April to August where the seasonal indices are greater than 100.

The following scatter plot and time series graph show the unseasonalised data plotted against time:





Overll, sales has increased over the period January 1998 (t=1) to December 2005 (t=96). The upper trend is combined with seasonal effects, and a close inspection of the scatter plot helps the detection of seasonality.

The seasonalised trend line for the Dofasco sales data is given below:

$y_t = 132.2754 + 4.838845t$ where $t = 1,2,3,.....96$.

The trend line with seasonalised sales does not give the correct view of sales behaviour because the seasonal effect is not included in the model. Instead, with a more appropriate decomposition model, one that includes both trend and seasonality, the trend line will look as follows (sales data has been deseasonalzed):

$y_t = 128.0653 + 4.947276t$

The above trend line used with the corresponding seasonal index will yield the forecasted monthly sales. For example, the forecasted sales for January 2006 will be:

$y_{97} = [128.0653 + 4.947276(97)][0.807978] = \$491.2 millions$.

2. Using Excel, several forecasting techniques were explored. Three moving average methods were used. These were unweighted moving averages of 4 years, 3 years, and 2 years producing MADs

of 3.9043, 2.6794, and 1.9004. The 2-year moving average produced the smallest MAD and seemed to be the best fit.

Next, single exponential smoothing models were examined with various values of α. For $\alpha = 0.3$, the MAD is 3.847. For $\alpha = 0.6$, the MAD is 5.7872. For $\alpha = 0.9$, the MAD is 13.251. The lower the value of α, the better the forecast. This means that the error has less impact on the new forecast, which is likely to be similar to the old one. At this point, we can clearly see that moving average forecasting produces a smaller error than exponential smoothing. This is attributable to the downward trend in the data, where high damping factors produce larger forecasting errors.

Trend analysis was performed on these data by fitting a line through the data. The Excel trend analysis resulted in MAD = 4.9418. A quadratic fit resulted in MAD = 2.8865. The model produced by this analysis is $y_t = 91.99912 - 6.35441t + 0.27929t^2$. The quadratic model reduces the estimating error much more efficiently than the linear trend model.

Overall, for Dofasco, the forecasting method that is the most effective in reducing error is the 2-year moving average, with MAD=1.9004. This is supported by a decreasing trend and as such, exponential smoothing is not as effective. On the other hand, a quadratic model is effective in reducing error, but not as much as a moving average method (2-year moving average in this case). With a bigger data base, a quadratic model might have been robust enough to reduce the error as effectively as the moving average method, if not more effective. This assertion can only be corroborated with more data, and under the present circumstances (a horizon from 1995 to 2008), a 2-year moving average is the recommended choice for forecasting the price-unit labour cost at Dofasco.

Chapter 16
Foot Locker in the Shoe Mix

1. Has the distribution of shoe sales by Price Category changed from the year 2005 to the year 2024? A chi-square goodness-of-fit test can be used to test this. Let the 2005 distribution be the expected frequencies and the 2024 values be the observed frequencies.
First, we work out this Chi-square Goodness-of-Fit test "long-hand" with the help of Excel:

The total of the 2024 observations is 297, but the total of the 2005 observations is only 291. So the first step is to determine the proportions of each year 2005 value as it is using the total of 291. Next multiply each of these proportions by 297 to get the expected values to compare to the year 2024. After that, compute the chi-square contribution of each pair and then sum the column to get the observed chi-square value.

2005	Expected Proportion	Expected Value	2024	chi-sq. contribution
115	.395189003	117.371134	126	0.634375
38	.130584192	38.783505	40	0.038157
37	.127147766	37.762887	35	0.202144
30	.103092784	30.618557	27	0.427648
22	.075601375	22.453608	20	0.268117
21	.072164948	21.432990	20	0.095808
11	.037800687	11.226804	11	0.004582
17	.058419244	17.350515	18	0.024312
			Observed χ^2 =	**1.695143**

Using the sample size of $n = 8$, df = 7, and $\alpha = .05$, a critical chi-square value of 14.0671 is attained. Since the observed chi-square is less than the critical chi-square, the decision is to fail to reject the null hypothesis. There is not enough evidence to declare that the 2024 distribution of shoe sales is any different than the 2005 distribution of shoe sales. Implications to Foot Locker and Nike is that the target market has not changed. This might mean that marketing attempts to change the target market have not been effective. It might also mean that the production schedule for various types of shoes need not change. Demand for shoes at the various price levels remains constant. If the companies desire to sell more high-end shoes, they need to make a renewed effort to effect changes in these distributions. This result basically tells them that they are where they were.

2. A chi-square test of independence is used to determine if Sex is independent of Location. Shown below is output for the analysis of this question.

Chi-Square Test: Male, Female

Expected counts are printed below observed counts

	Male	Female	Total
U.S. West	29	43	72
	41.27	30.73	
U.S. South	48	20	68
	38.98	29.02	
U.S. East	52	61	113
	64.77	48.23	
U.S. North	28	25	53
	30.38	22.62	
Europe	78	32	110
	63.05	46.95	
Australia	47	29	76
	43.56	32.44	
Total	282	210	492

Chi-Sq = 3.647 + 4.898 + 2.090 + 2.806 + 2.517 + 3.380 +
0.186 + 0.250 + 3.545 + 4.761 + 0.272 + 0.365 = **28.716**
DF = 5, P-Value = **0.000**

An observed chi-square of 28.716 was obtained with an associated *p*-value of .000 on this question. This indicates that sex is not independent of location when it comes to the number of formal suggestions. A cursory examination of the observed values reveals that more suggestions were submitted by females than males in the U.S. West and the U.S. East. However, more suggestions were submitted by males than females in all other regions. In the U.S. South and in Europe the ratio of formal suggestions of males to females was over two to one. What this result says is when it comes to making suggestions, the regional culture has an impact on who makes the suggestions. Foot Locker might make a concerted effort to increase participation by the sex with fewer suggestions in each region and perhaps make an attempt to change corporate culture so that regional tendencies do not affect the employee suggestion program.

Chapter 17
Schwinn

1. Since two independent samples are being compared and the shape of the population distribution is unknown, the Mann-Whitney *U* test is used to analyze the data rather than the *t* test for two independent samples. The null hypothesis is that there is no difference between the age of Schwinn customers in Colorado Springs and the age of Schwinn customers in Alberta and British Columbia. The Minitab computer output for the Mann-Whitney test is shown below. The analysis reveals that the median age for a customer in Colorado Springs was 31 years of age and the median age for a customer in Alberta and British Columbiais 14 years. It appears that the target market in Colorado Springs is young adult versus Alberta and British Columbia where it is children and early teens. Since one of the sample sizes is greater than 10, a *large* sample Mann-Whitney *U* test is the appropriate test. As noted in the text, the Minitab Mann-Whitney test does not yield a *z* statistic but rather gives the value of *W* and a *p*-value. Since the *p*-value is .0030, the null hypothesis is rejected at $\alpha = .01$. There is a significant difference between the age of Schwinn customers in Colorado Springs and those in Alberta and British Columbia. The median ages support the theory that a much older group of customers purchase Schwinn bikes in Colorado Springs perhaps to be used in mountain biking. The St. Louis population appears to be a youth market.

Mann-Whitney: Colorado Springs, Alberta & British Columbia

Method

η_1: median of Colorado Springs
η_2: median of Alberta and British Columbia
Difference: $\eta_1 - \eta_2$

Descriptive Statistics

Sample	N	Median
Colorado Springs	11	31
Alberta and British Columbia	9	14

Estimation for Difference

Difference	CI for Difference	Achieved Confidence
17	(9, 23)	95.18%

Test

Null hypothesis: $H_0: \eta_1 - \eta_2 = 0$
Alternative hypothesis: $H_1: \eta_1 - \eta_2 \neq 0$

Method	W-Value	P-Value
Not adjusted for ties	155.00	0.003
Adjusted for ties	155.00	0.003

2. Since three independent groups are being compared, it would appear that this is a completely randomized design and that a one-way ANOVA could be used to analyze the data. However, it is uncertain whether the data are normally distributed or not, so a Kruskal-Wallis test is used to analyze the data. The independent variable is supplier with three classifications: supplier 1, supplier 2, and supplier 3. The dependent variable is the weight of the handle bar. The null hypothesis is that there is no difference in the weight of handle bars supplied by the three suppliers. The alternative hypothesis is that there is a difference in the weights of handle bars by supplier. Minitab was used to analyze these data. The results, given below, show that there is no significant difference in the weights of handle bars according to supplier. The *H* value (3.09) is Minitab's equivalent to the Kruskal-Wallis *K* value. The associated *p*-value (.213) denotes that there is no significant difference even at $\alpha = .10$. The differences in the medians, shown in the Minitab output, are merely due to chance. To Schwinn this might mean that, at least on handle bar weight, these suppliers are interchangeable.

Kruskal-Wallis Test: Weight versus Supplier

Descriptive Statistics

Supplier	N	Median	Mean Rank	Z-Value
1	8	201.935	12.4	1.57
2	6	198.160	9.5	-0.26
3	5	195.180	6.8	-1.48
Overall	19		10.0	

Test

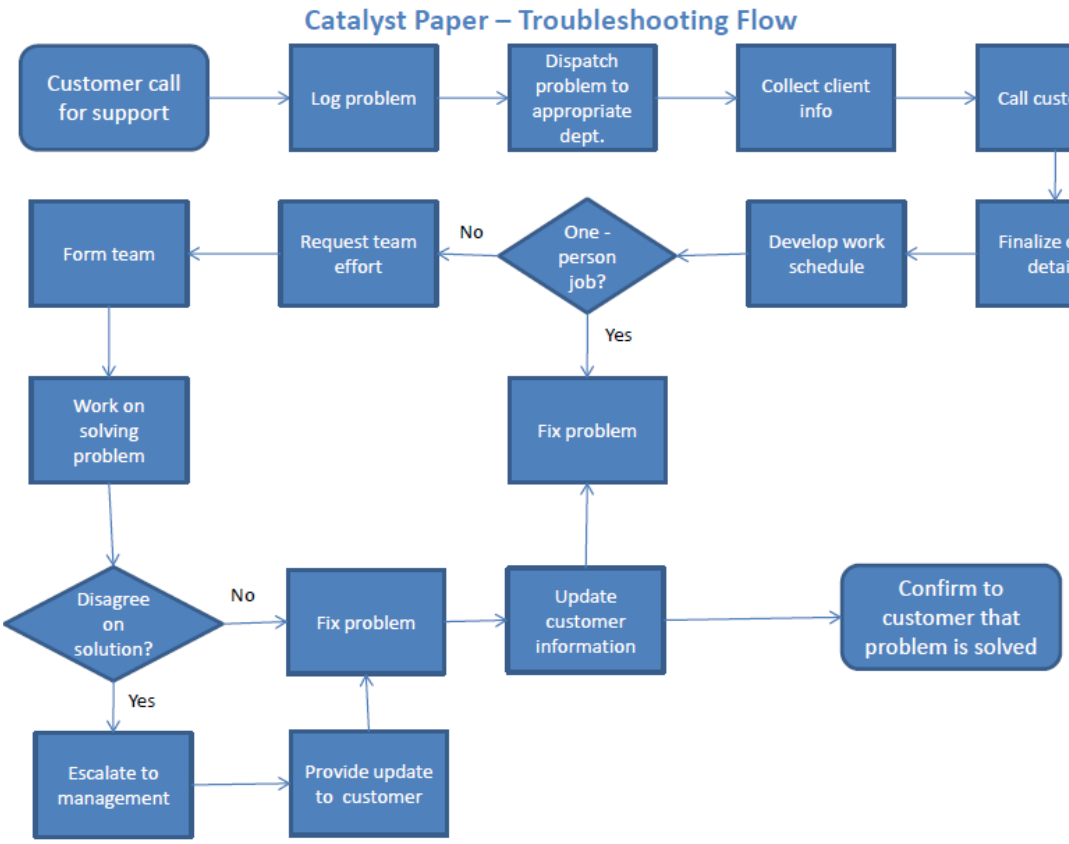
Null hypothesis H_0: All medians are equal
Alternative hypothesis H_1: At least one median is different

DF	H-Value	P-Value
2	3.09	0.213

3. The Minitab output for the Runs Test contains a *p*-value of .9083 that is not significant (underscored by Minitab's statement – Cannot reject at alpha – 0.05). This indicates that the paint flaws are occurring in a random manner. In this sample, there are 19 observations above *K* and 56 below (out of 75 bicycles). Since the data were coded with a 1 to indicate at least one flaw and a 0 to indicate no flaws, there are 19 bicycles with at least one flaw out of the 75 bicycles. The proportion of flawed bicycles is $\frac{19}{75} = .2533$ (shown as *K* in the output). While Schwinn management may be happy to know that there appears to be no systematic pattern of flaws, they are likely to be unhappy about having at least one flaw in 25.33% of the bikes. Perhaps, an intense, manufacturer-wide quality effort could be made to reduce the percentages of flaws. After, studying chapter 18 (Statistical Quality Control), the student might be able to recommend the use of a Pareto Chart to prioritize the types of flaws that are occurring and a Fishbone diagram to assist managers and production engineers in searching for causes of the flaws which might include such things as raw materials, machinery, technique, and workers.

Chapter 18
Catalyst Paper Introduces Microsoft Dynamics CRM

1.



2. The variables that are most likely be used by catalyst to improve quality are many fold. Among the important ones, we have: time reduction in handling customer requests; increased productivity expressed in reduced time to perform tasks; percentage of errors in the operations of the CRM system; number of clients served by the customer service representative; accuracy of pricing and quotes. Quality characteristics that should be taken into consideration. The quality characteristics that should be kept in mind for the manufacturing and sales processes for catalyst might be: a) for manufacturing: performance, reliability, durability of the product manufactured, and b) for sales: the speed to which the product is delivered, the price of the product, lead time to deliver to the customer, and the responsiveness to customer needs and demands. All three types of control charts can be used to monitor whether the manufacturing and sales processes are in control or not. The variables discussed are either continuous or discrete. In the former case, x-bar charts and R-charts are used together to determine the pattern of variation of the collected data on the variables. If the data is discrete, then either a p-chart will be used if the data is expressed in binary form (success/failure, i.e., percentage of errors) or a c-chart

if the data is expressed in rates (value per unit, i.e., number of clients served by customer service representative). Patterns of variation must be carefully analyzed to ensure that special cause variation is eliminated if it is present. In doing so, the process will be stabilized (in control). In order to ensure that the quality monitoring process is well managed, data collection methods must be free of any bias, either in selecting the samples or in collecting the data from them.

3. Control charts are an important aspect and tool of the quality control process and overall management of quality. However, there are other quality tools which are very useful in driving the quality initiative effectively. Catalyst should consider the use of **process flow diagrams** to help in the identification of possible points in the process where problems occur; **Pareto charts** are also useful in helping catalyst focus on the most important problems; and Cause-and-effect diagrams (also known as "Fishbone diagrams") which help in the approach to seek the possible causes of a problem. These three tools, and other secondary ones such as checklists, histograms and scatter plots will help the overall problem solving and continuous improvement objectives for Catalyst. These tools combined with the appropriate use of control charts help management address the entire quality initiative, which includes after-process inspection.

Chapter 19
Fletcher-Terry: On the Cutting Edge

1. There are several decisions that management had to make during this time including 1) whether or not to invest in technology, 2) expand its line of offerings through imports, 3) conduct a significant planning process, 4) attempt to increase market share, 5) develop new products, 6)create greater employee involvement, 7) invest in employee education, 8) invest in plant improvements, and 8) implement a participatory management system.

   Several states of nature occurred which could have affected Fletcher-Terry's outcomes. Some of these include: 1) its largest customers decided to introduce their own private-label cutters made overseas, 2) the technology that Fletcher-Terry invested in would not work, 3) dollar weakened, 4) slow-down in demand for cutters, 5) employees do not respond to company efforts.

2.

		$ Up (.25)	$ Same (.35)	$ Down (.40)
Decision Alternative	Import	$350,000	$275,000	-$555,000
	Not Import	-$22,700	-$22,700	-$22,000

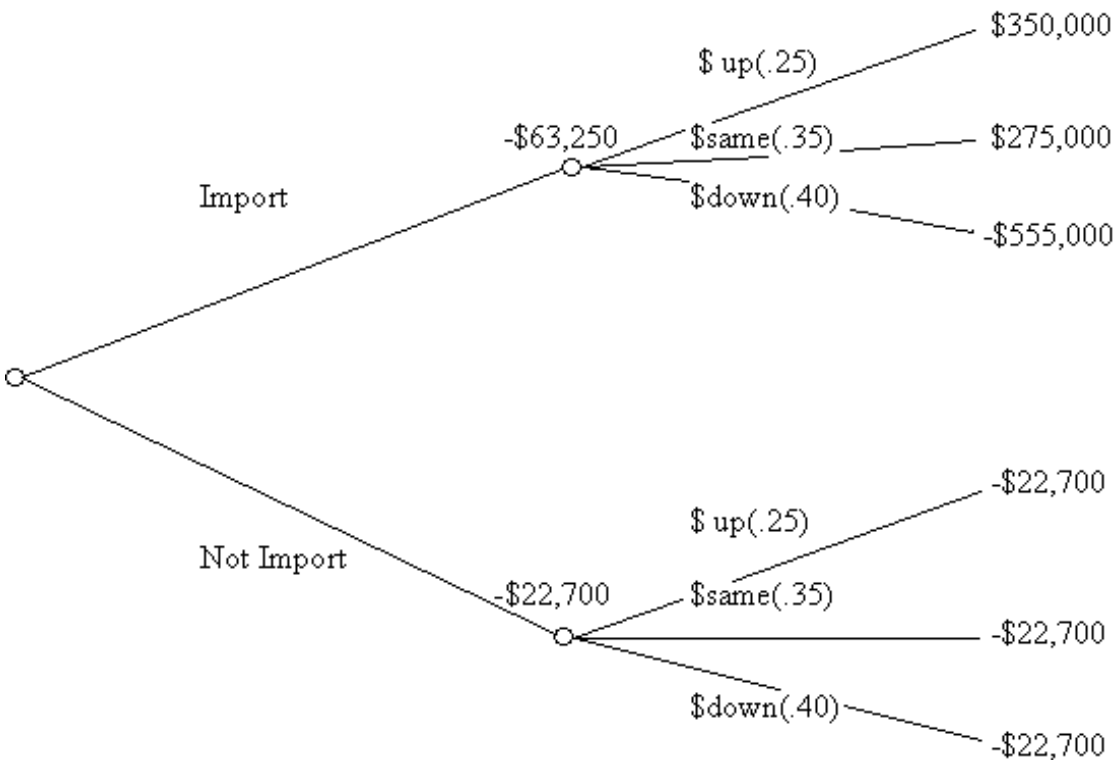
   The expected monetary value (EMV) for each of these alternatives is:

   Import: (.25)($350,000) + (.35)($350,000) + (.40)(-$555,000) = -$63,000

   Not Import: (.25)(-$22,700) + (.35)(-$22,700) + (.40)(-$22,700) = -$22,700

   The EMV'er would choose the highest of these alternatives which is to not import and take a $22,700 loss. The risk avoider would also choose to not import. However, a risk taker might decide to import gambling that the dollar does not go down.

The decision table and the expected values are displayed below in a decision tree.



Business Statistics, Canadian Edition
Database Exercise Answers

Chapter 1 Introduction to Statistics and Business Analytics

Question 1:

a. **Answer:** Nominal
b. **Answer:** Interval
c. **Answer:** Interval