Module 3: Quantitative Decision Tools

Module Three: Quantitative Decision Tools

3.01 Learning Objectives

Learning Objectives

After completing this module, you should be able to:

1. Evaluate the usefulness of different statistical techniques and their real-world application
2. Describe the various types of regression analysis and their real-world application
3. Analyze the results of a regression analysis
4. Describe common problems with multiple regression
5. Describe other statistical techniques and their real-world application
6. Explain the advantages and disadvantages of various statistical techniques
7. Choose a statistical technique based on a brief case study

3.03 Linear Programming

Linear Programming

Linear programming is a mathematical technique used to find a maximum or minimum of linear equations containing several variables. Problems in business often involve decisions about how best to use limited resources (for example time, money, space) in the most efficient manner. Typically this involves determining how to allocate the resources such that costs are minimized, or profits are maximized.

Consider this sample business question: "How much of each product should we ship from each warehouse to our various retail locations?"

In this problem, there is a physical limitation on the amount of merchandise a truck can carry from one warehouse to the stores on its route. Further, each warehouse stores different products. There may be time or cost considerations based on the mode of transportation or quality considerations based on the type of product being shipped. These are examples of the restrictions, or constraints, the decision maker must work around.

There is also an objective the decision maker considers when deciding which course of action is best. Are we trying to minimize total transportation costs? Are we trying to utilize fewer trucks per day?

Computers and spreadsheet programs allow managers to use linear programming easily and efficiently to help make decisions. Computers today can easily carry out a method known as the simplex method, a complicated mathematical method that helps solve linear programming problems.

Once we know how to express problems as linear programming models, we can use the solution found in a program such as Excel. Here, we will focus on constructing linear programming models.

Application: The product mix problem

Managers must often find the optimal mix of products to maximize profit. Here is an example of a product mix problem.

MindSledge produces three types of Sledgehammers. We label the three types of sledgehammers 1, 2 and 3. These sledgehammers differ from each other in size, length, and materials used. MindSledge sells each of the three products for a different price. Next week, MindSledge has the following maximum capacity at the following costs:

2,000 hours of labor	$8.00 per hour
3,000 lbs of metal	$.50 per lb
5,000 lbs of wood	$1.00 per lb

In other words, the total number of labor hours used to produce all three sledgehammer types must be less than or equal to the total number of labor hours available, which is 2,000.

Also, the market dictates that it is impossible to sell more than 700 type 1 sledgehammers, 1100 type 2 sledgehammers, and 1300 type 3 sledgehammers within a week. MindSledge wants to maximize their weekly profit.

	Sledge 1	**Sledge 2**	**Sledge 3**
Labor (hours)	1	2	3
Metal (pounds)	3	1	2
Wood (pounds)	2	3	1
Selling Price	$64.00	$81.00	$100.00

The first step in formulating a linear programming model is to understand the problem. In this case, the problem is to figure out how many of each type of sledgehammer to produce to maximize profit, while using no more than 2,000 hours of skilled labor, 3,000 pounds of metal and 5,000 pounds of wood.

We can calculate the profit on each sledgehammer sold:

selling price-labor - materials = profit

$52.50

	Sledge 1	**Sledge 2**	**Sledge 3**
Selling Price	$64.00	$81.00	$100.00
Labor	$8.00	$16.00	$24.00
Metal	$1.50	$0.50	$1.00
Wood	$2.00	$3.00	$1.00

Profit	$61.50	$74.00	

Decision-makers must also identify the restrictions on the amount of each sledgehammer type that can be produced. These are the constraints in the linear programming model. Recall the three constraints facing MindSledge.

1. 2,000 hours of labor are available
2. 3,000 pounds of metal are available
3. 5,000 pounds of wood are available

MindSledge also faces three "upper bound" constraints on the decision variables:

1. Maximum demand for Sledge 1 is 700 units
2. Maximum demand for Sledge 2 is 1,100 units
3. Maximum demand for Sledge 3 is 1,300 units

A manager's goal is to determine how many of each sledgehammer sold (without exceeding the demand) maximizes profit while simultaneously satisfying the constraints listed above. The constraints define the set of feasible solutions for the problem.

While it is possible to solve this problem using mathematical equations, this becomes complicated when there are multiple decision variables. Fortunately, these mathematical techniques are built into spreadsheet packages that make solving linear programming problems relatively easy.

3.04 Crossover Analysis

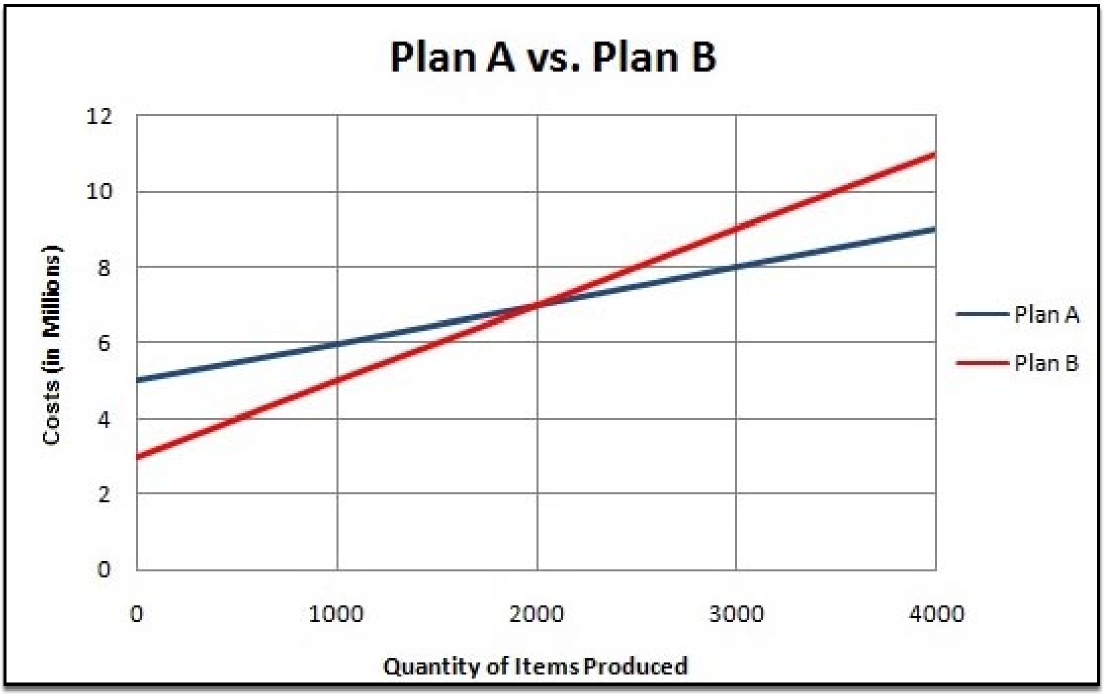
Crossover Analysis

When there are two or more plans or options to consider, crossover analysis allows a decision maker to identify the crossover point, which represents the point at which we are indifferent between the plans. With the crossover point identified, it also clarifies which option is better on either side of the crossover point.

For example, let's assume that our objective is to minimize cost.

Plan A	high fixed costs	low variable costs
Plan B	low fixed costs	high variable costs

If we compiled a table comparing costs for the two plans at each number of units sold, we would find a specific number of units at which both strategies have the same costs. This represents the crossover point.

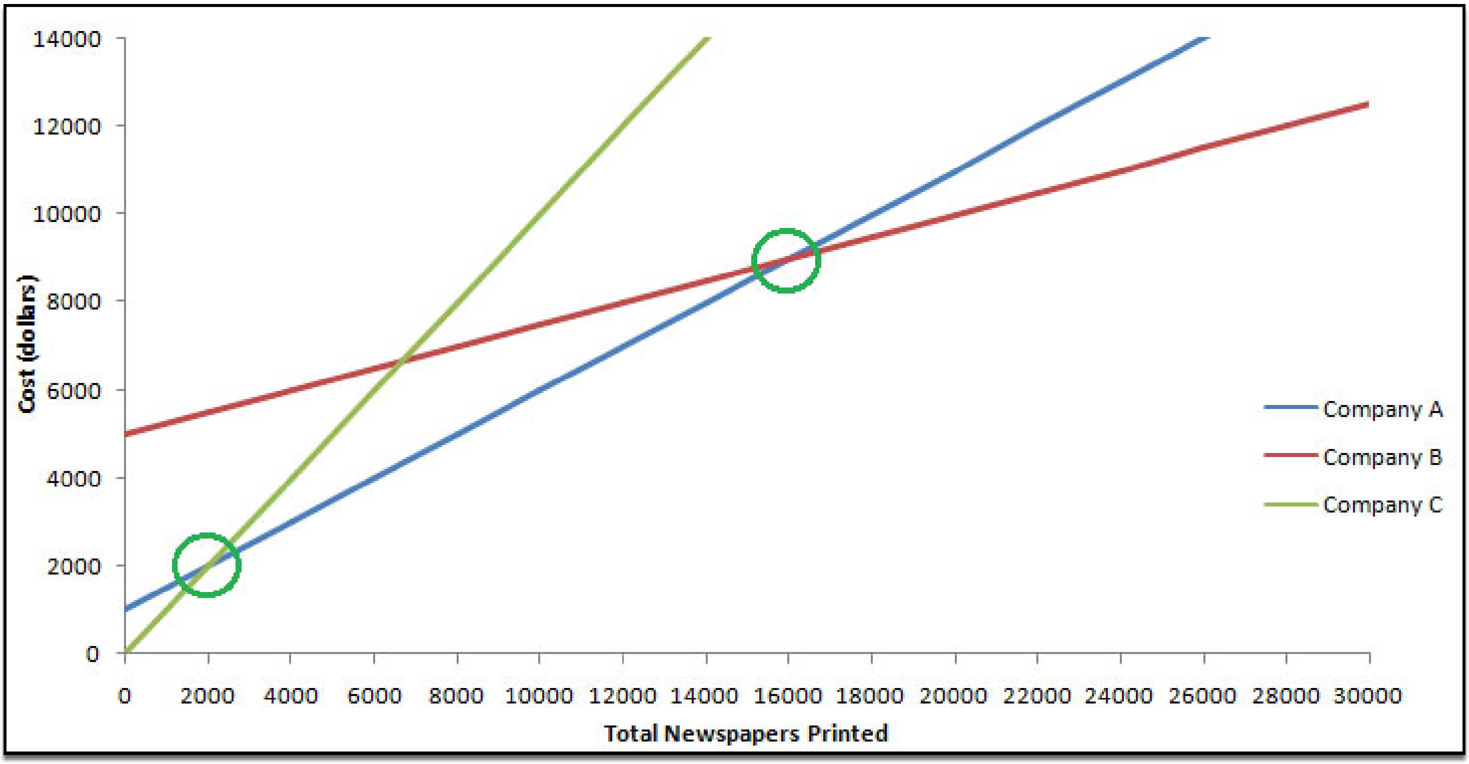


A decision maker who is interested in minimizing costs can now apply this analysis to his or her situation. This graph shows that if the company projects unit sales under 2,000, it makes sense to pursue Plan B which achieves lower costs at that level of sales. If, however, the company projects sales above 2,000, it makes sense to pursue Plan A which achieves lower costs at a larger sales volume.

Application: Real-world Plan A vs. Plan B vs. Plan C

A small-town newspaper is considering outsourcing printing. We are trying to decide which printing vendor to use. Printing Company A costs $1,000 monthly and $.50 to print each newspaper. Printing Company B costs $5,000 monthly and $.25 to print each newspaper. Printing Company C has no fixed cost but costs $1 to print each copy of the newspaper.

We can graph the three lines as follows:



Our goal is to minimize our cost of printing. Analyzing the graph, we see relevant crossover points circled. To the left of the first crossover point, we see that if we print fewer than about 2,000 newspapers, Company C is the cheapest printing vendor. Between the two crossover points, we can see that it is the lowest cost would be to use Company A. To the right of the second crossover point, we see that Company B is the least expensive vendor.

Algebraic Equation of Crossover Analysis

Here are the equations for each company's cost to print per newspaper:

Printing Company A: $y = 0.5x + 1000$

Printing Company B: $y = 0.25x + 5000$

Printing Company C: $y = 1x$

Using the equations for each of the lines, we can algebraically find these exact crossover points. By setting the equations for Company A and Company C equal to each other and solving for x, which represents the total newspapers printed, we find that this crossover point between Company C and Company A occurs at 2000 newspapers. Similarly, we find algebraically that the crossover point between Company A and Company B occurs at 16,000 newspapers. Therefore, if we plan on printing fewer than 2000 newspapers next month, we will choose Company C to print our newspapers. If we print between 2000 and 16,000 newspaper, we should choose Company A, and if we plan on printing more than 16,000 newspapers, Company B is the wise choice of vendor.

3.05 Break-even Analysis

Break-even Analysis - Example Using the Equation Method

Consider a small company, BG Manufacturing, which produces and sells ceiling fans. BG Manufacturing has a ceiling fan which is priced competitively at $49 a unit. Management knows that production costs are $24 per unit and that overhead (salaries, rent, equipment depreciation, etc.) is $75,000. What volume will BG Manufacturing need to sell to break even?

Based on recent market trends, management expects to sell from 1,000 to 5,000 units. Based on these numbers, the following spreadsheet was developed:

A	B	*C = A × B*	D	*E = A × D*	*C - E*	F	*C - E - F*
Units Sold	**Market Price per Unit**	**Sales**	**Variable Cost per Unit**	**Variable Costs**	**Contribution Margin**	**Fixed Costs**	**Pretax Profit**
0	$49	$0	$24	$0	$0	$75,000	($75,000)
500	$49	$24,500	$24	$12,000	$12,500	$75,000	($62,500)
1,000	$49	$49,000	$24	$24,000	$25,000	$75,000	($50,000)
1,500	$49	$73,500	$24	$36,000	$37,500	$75,000	($37,500)
2,000	$49	$98,000	$24	$48,000	$50,000	$75,000	($25,000)
2,500	$49	$122,500	$24	$60,000	$62,500	$75,000	($12,500)
3,000	**$49**	**$147,000**	**$24**	**$72,000**	**$75,000**	**$75,000**	**$0**
3,500	$49	$171,500	$24	$84,000	$87,500	$75,000	$12,500
4,000	$49	$196,000	$24	$96,000	$100,000	$75,000	$25,000
4,500	$49	$220,500	$24	$108,000	$112,500	$75,000	$37,500
5,000	$49	$245,000	$24	$120,000	$125,000	$75,000	$50,000

As the table indicates, break even is achieved at 3,000 units. For any ceiling fans sold beyond 3,000, BG Manufacturing is making a profit (prior to taxes); if it sells less than 3,000 it loses money.

The **margin of safety** is the excess of budgeted or actual sales over the break-even volume of sales (the numbers in green in the chart above.)

Calculating the break-even volume is straightforward using the break-even formula:

$$\text{Break-even Units} = \frac{\text{Fixed Costs}}{\text{Contribution Margin per Unit}}$$

$$= \frac{\text{Fixed Costs}}{\text{Price} - \text{Variable Cost per Unit}}$$

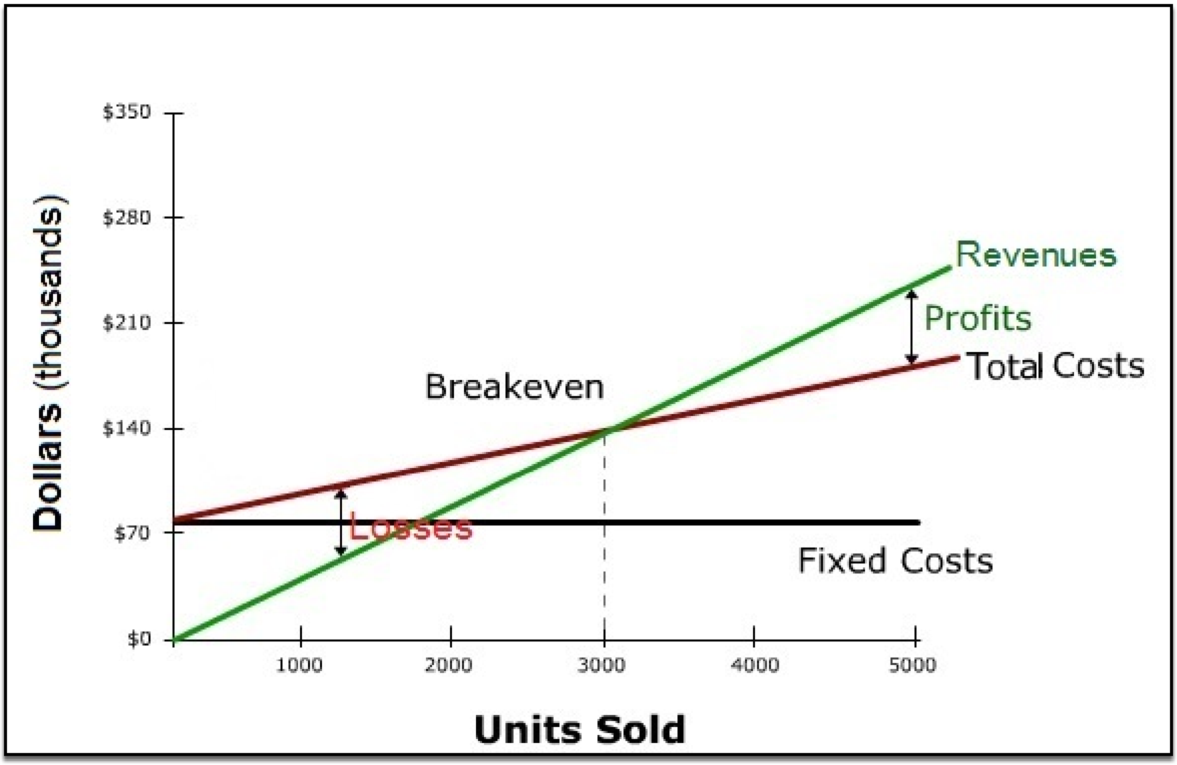
$$= \frac{\$75{,}000}{(\$49-\$24)}$$

$$= \frac{\$75{,}000}{\$25}$$

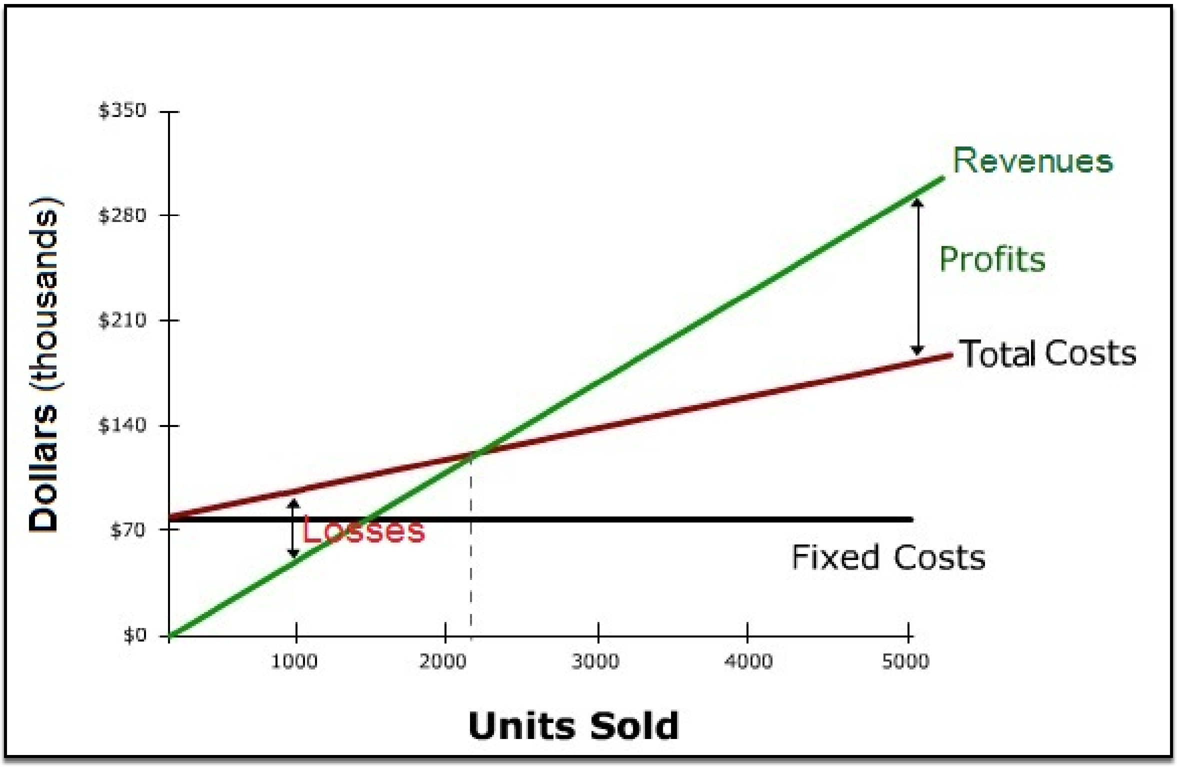
= 3,000 Units (fans)

Break-even analysis - Example using Graphical Method

According to this graph, you can see that break-even occurs at 3,000 units sold when the unit price is $49.



What will happen if the sale price is increased from $49 to $59?



If the price per unit is increased from $49 to $59, the amount of sales needed to cover the total costs decreases. Notice that the sales line moves from $245,000 at 5,000 units to $295,000 at 5,000 units. Also notice that, due to the increase in sales revenue, while costs remain the same, the break-even point goes down from 3,000 to 2,143.

3.06 Hypothesis Testing

Hypothesis Testing

Hypothesis testing is the method of inferential statistics used to make decisions or judgments about population parameters. A common type of hypothesis is a statement or claim about a given population.

To test a hypothesis, you must convert the question into a null hypothesis and alternative hypothesis.

Null Hypothesis

The null hypothesis, or H_0, is the statement that there is **no relationship.** For whatever relationship is being tested, the null hypothesis is the statement that the relationship does not exist. For example, if a test is conducted to determine a difference between two means, the null hypothesis will state that there is **no difference** between these two means.

Alternative Hypothesis

The alternative hypothesis, or H_A, is the opposite statement to the null hypothesis. It states that there **is a relationship** for whatever relationship is being tested. If we conduct a test to determine the difference between two means, the alternative hypothesis states that there **is a difference** between these two means.

Statistical Significance

In statistics, it is not enough to simply look at two numbers and proclaim that they are "different" from one another. Rather, we want to determine if a difference is statistically significant. A *statistically significant* result is unlikely to be caused by random variation or errors. A difference that is statistically significant is called a significant difference. A hypothesis test will tell us whether the results are statistically significant or not.

Writing Null and Alternative Hypotheses

Hypothesis testing can be used in many contexts. For example, you can execute a hypothesis test to determine whether two means, denoted μ_1 and μ_2, are *significantly different* from one another.

So, the null hypothesis statement can be written as:

$$\mathrm{H}_0 : \mu_1 = \mu_2$$

The alternative hypothesis can take the following form:

$$\mathrm{H}_A : \mu_1 \neq \mu_2$$

The null hypothesis is the statement that is being tested. There are two possibilities after conducting a hypothesis test:

- Reject the null hypothesis.
- Fail to reject the null hypothesis.

Notice that both of these possibilities pertain to the null hypothesis. The null hypothesis is always the statement that is being tested. The outcome of your experiment is to determine whether the null hypothesis should be rejected. If you reject the null hypothesis, the difference being tested is significant: the difference is most likely not caused by random variation or error. On the other hand, if you fail to reject the null hypothesis, you did not find a significant difference.

Step 1: State the Null Hypothesis and the Alternative Hypothesis

As we've seen:

- The null hypothesis, or H_0, is the statement that there is **no relationship.**
- The alternative hypothesis, or H_A, is the statement that there **is a relationship** for whatever relationship is being tested.

Imagine you are testing the effect of temperature on the growth of a species of orchids. To do this, the heights of mature orchids are measured. Two samples of orchids are being compared. One sample of orchids was grown in a greenhouse at a constant temperature of 65°F. The other sample of orchids was grown in a greenhouse at a constant temperature of 75°F. The mean height of orchids in these two groups are μ_1 and μ_2, respectively.

The null hypothesis, or H_0, is the statement that "there is no significant difference between the two means, μ_1 and μ_2." In other words, the null hypothesis states that you find **no relationship** between temperature and the height of the orchids.

$$H_0: \mu_1 = \mu_2$$

To test the validity of the null hypothesis, you must specify an alternative hypothesis, H_A.

Your alternative hypothesis would be that "there is a significant difference between the two means, μ_1 and μ_2." In other words, the alternative hypothesis states that you find there **is a relationship** between temperature and the height of the orchids. This hypothesis can be written as:

$$H_A: \mu_1 \neq \mu_2$$

Step 2: Decide on the Significance Level

The significance level is a decision criterion that specifies the degree of certainty with which you want to make your judgment of whether or not to reject the null hypothesis. The significance level is the probability that you will mistakenly reject a true null hypothesis based on the sample statistic. The more careful you want to be about not rejecting a true null hypothesis, the smaller your significance level should be:

- A **higher significance level** indicates a **higher threshold to reject the null hypothesis**. To state that there is a significant difference, you have to be more certain that random chance or error is not causing the difference.

- A **lower significance level** indicates a **lower threshold to reject the null hypothesis**. To state that there is a significant difference, you do not have to be as certain that random chance or error is not causing the difference.

A commonly used significance level in many research settings is 0.05. A 0.05 significance level means that you state that the results were significant if there is only a 5% chance it was actually caused by random variation or errors. You expect a rejected null hypothesis to be an incorrect decision in only 5% of cases. We will use a 0.05 significance level in most of our examples.

Another way of saying this is that if you find a significant result (i.e., your obtained p-value is less than .05) then you are 95% confident that you are correct in your decision (i.e.,your obtained result is within the **95% confidence interval**).

Step 3: Compute the Value of the Test Statistic

There are a few possible test statistics. Be sure to choose an appropriate test statistic for your hypothesis test.

For example, you may be testing the difference between a sample mean, and a population mean. One of the most useful kinds of test statistics (that can be used for hypothesis testing in this situation) is the One-Sample *t*-Test. The One-Sample *t*-Test can be used to test a null hypothesis concerning a population mean based on statistics from *one random sample* from the population.

The test statistic for a One-Sample *t*-test is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

$\bar{x}$ = the sample average
s = the standard deviation of the sample values
n = the number of values in the sample
$\frac{s}{\sqrt{n}} = S_{\bar{x}}$ = standard error of the mean

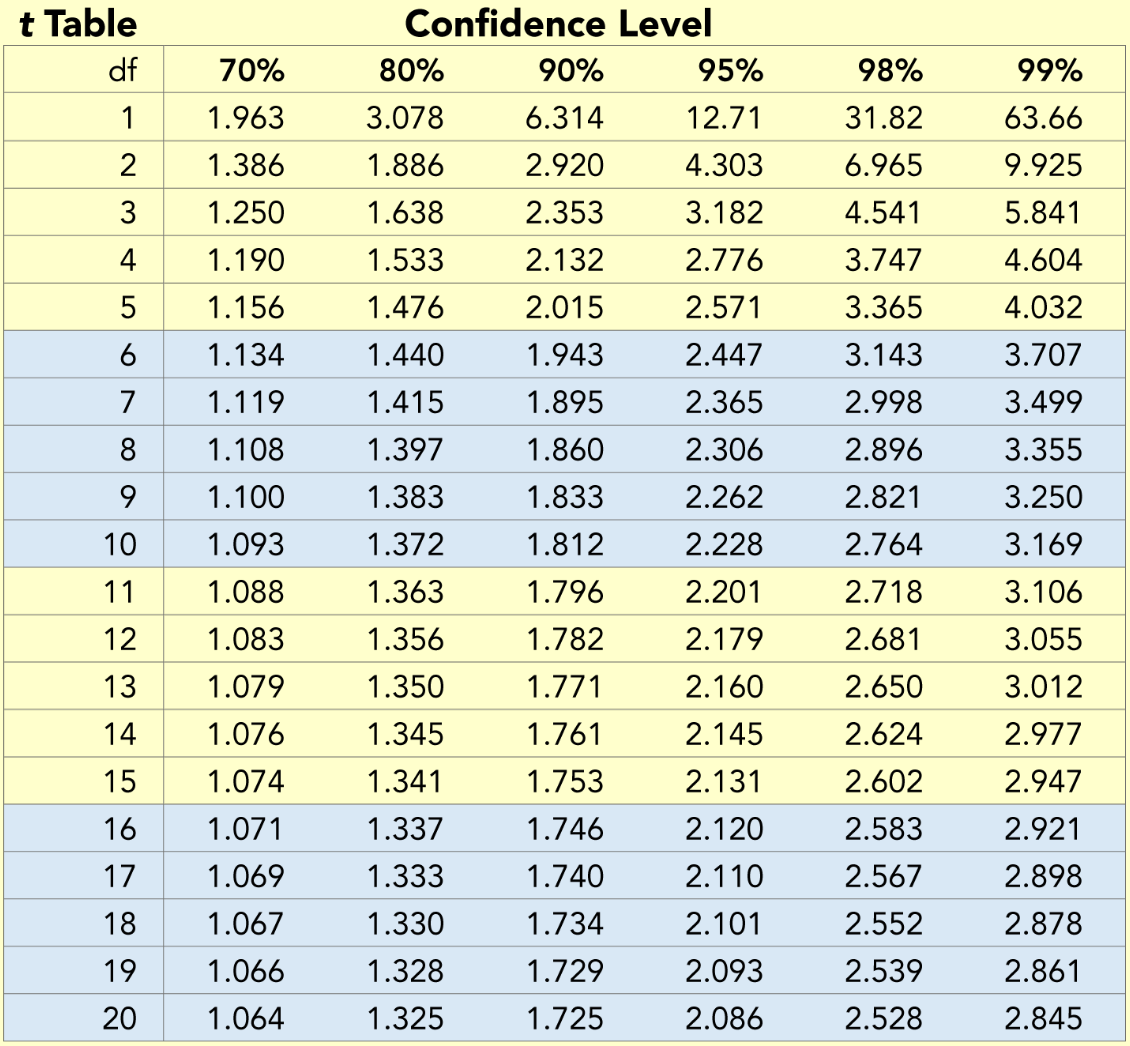
To compute the value of the test-statistic, we calculate or identify each of the necessary values $\bar{x}$, s, n, μ_0. After plugging the appropriate values into the formula, the test statistic value (*t*) is computed using arithmetic. The value of the test statistic is crucial as we move on to the fourth and final step of our hypothesis test.

Step 4: Find the Critical Value and Compare to Test Statistic Value

Ultimately, the goal of the hypothesis test is to make a decision about the null hypothesis. This determination is made by comparing the critical value to the test statistic value.

The critical value depends on upon the significance level and the test statistic that was employed. Once we calculate the test statistic value, the critical value needs to be found. Having both of these values, we can compare them to one another. If the absolute value of the test statistic is greater than the critical value, we reject the null hypothesis: there is statistical significance. So, the critical value is the tipping point between where we reject the null hypothesis and where we fail to reject the null hypothesis.

To determine whether the test statistic's absolute value is large enough to reject the null hypothesis, we must find the critical values for the hypothesis test from a distribution table. A distribution table corresponds to the test statistic used. The table for a t-test, known as a t table, is partially shown in the image below. The critical value within this table depends upon a couple of factors: degrees of freedom and the significance level or confidence level.



***t* Table** **Confidence Level**

df	70%	80%	90%	95%	98%	99%
1	1.963	3.078	6.314	12.71	31.82	63.66
2	1.386	1.886	2.920	4.303	6.965	9.925
3	1.250	1.638	2.353	3.182	4.541	5.841
4	1.190	1.533	2.132	2.776	3.747	4.604
5	1.156	1.476	2.015	2.571	3.365	4.032
6	1.134	1.440	1.943	2.447	3.143	3.707
7	1.119	1.415	1.895	2.365	2.998	3.499
8	1.108	1.397	1.860	2.306	2.896	3.355
9	1.100	1.383	1.833	2.262	2.821	3.250
10	1.093	1.372	1.812	2.228	2.764	3.169
11	1.088	1.363	1.796	2.201	2.718	3.106
12	1.083	1.356	1.782	2.179	2.681	3.055
13	1.079	1.350	1.771	2.160	2.650	3.012
14	1.076	1.345	1.761	2.145	2.624	2.977
15	1.074	1.341	1.753	2.131	2.602	2.947
16	1.071	1.337	1.746	2.120	2.583	2.921
17	1.069	1.333	1.740	2.110	2.567	2.898
18	1.067	1.330	1.734	2.101	2.552	2.878
19	1.066	1.328	1.729	2.093	2.539	2.861
20	1.064	1.325	1.725	2.086	2.528	2.845

To find the critical value, you'll need to calculate the degrees of freedom (df) for the test, which is the sample size minus one (*n* - 1)).

We determine your critical t-test statistic's value by finding the table value in the row corresponding to the degrees of freedom (df), and proper column corresponding to the significance level or confidence level. From the t-table, we locate and identify the critical value.

There are two possible outcomes:

- **If the test statistic's value is NOT greater than the critical value, we FAIL TO REJECT the null hypothesis** For a *t*-test, this means is that the difference between the sample mean and μ_0 is not significant. There is no significant difference. There is not enough evidence to conclude that the null hypothesis is false.

- On the other hand, **if the test statistic's value IS greater than the critical value, we REJECT the null hypothesis**. For a *t*-test, this means is that the difference between the sample mean and μ_0 is significant. There is a significant difference. Therefore, there is enough evidence to conclude that the null hypothesis is false.

Hypothesis Testing Practice

The concept of hypothesis testing is very difficult to conceptualize. The only way to understand and master hypothesis testing is through practice. The following pages have examples of hypothesis tests. Be sure to carefully examine these problems to see hypothesis testing in action!

3.06.1 Hypothesis Testing Example, One-Sample *t*-Test

Hypothesis Testing Example, One-Sample *t*-Test

Example: Fitness Club Visits

Imagine that you are working for a large chain of fitness clubs. Five years ago, your customers worked out at the club an average of 2.3 days per week. You suspect that customers today are exercising more frequently, so you are interested in whether or not the null hypothesis — "the average number of times a customer visits the fitness club in a week is 2.3" — is still true. With this as your null hypothesis, your alternative hypothesis would then be written as "the average number of times a customer visits the fitness club in a week is not equal to 2.3."

You survey a random sample of the club's customers and ask them each how many times they visit the fitness club each week. The results from your survey are listed in the following table:

Customer Survey Responses: Number of Visits to Fitness Club in a Week (*n* = 30)

0	3	3	0	3	4
1	2	3	2	3	4
2	2	2	1	3	3
1	2	3	2	3	4
1	2	5	2	2	6

You calculate the sample mean for the data in this table as 2.47. Follow the steps below, and use the following sample data as inputs for your hypothesis testing, to determine whether this difference is significant.

Hypothesis Testing Steps, Fitness Club Example

- Step 1
- Step 2
- Step 3
- Step 4

Step 1: State the Null Hypothesis and the Alternative Hypothesis

For our example, the null hypothesis is: "the average number of times a customer visits the fitness club in a week is 2.3." The null hypothesis states that there is **no significant difference** between the sample mean and the known value of 2.3. The null hypothesis can be shown as:

$$H_0: \mu = 2.3$$

To test the validity of the null hypothesis, you must specify an alternative hypothesis, H_A.

Your alternative hypothesis is: "the average number of times a customer visits the fitness club in a week is not equal to 2.3." The alternative hypothesis states that there **is a significant difference** between the sample mean and the known value of 2.3. The alternative hypothesis can be written as:

$$H_A: \mu \neq 2.3$$

Step 2: Decide on the Significance Level

The significance level is a decision criterion that specifies the degree of certainty with which you want to make your judgment of whether or not to reject the null hypothesis. The significance level is the probability that you will mistakenly reject a true null hypothesis based on the sample statistic. The more careful you want to be about not rejecting a true null hypothesis, the smaller your significance level should be:

- A **higher significance level** indicates a **higher threshold to reject the null hypothesis**. To state that there is a significant difference, you have to be more certain that random chance or error is not causing the difference.

- A **lower significance level** indicates a **lower threshold to reject the null hypothesis**. To state that there is a significant difference, you do not have to be as certain that random chance or error is not causing the difference.

A commonly used significance level in many research settings is 0.05. A 0.05 significance level means that you state that the results were significant if there is only a 5% chance it was actually caused by random variation or errors. You expect a rejected null hypothesis to be an incorrect decision in only 5% of cases. We will use a 0.05 significance level in this example.

Step 3: Compute the Value of the Test Statistic

In this example, we are testing a difference between means. Our hypothesis test will allow us to determine if a difference between the sample mean and the population mean is significant. For the fitness club example, the population mean is 2.3. Therefore, the farther the sample mean is from 2.3 (in addition to other factors such as sample size and variance), the more reason we have to conclude that the null hypothesis is false.

To choose the best test statistic, we need to examine different aspects of our example, such as the sample size. A*t*-test is suitable in this situation. Here, we can use a One-Sample t-Test. As the name indicates, a *One Sample* *t*-Test is used when the mean of *one sample* is being tested against a known value. In this example, we are testing the sample mean against the population mean, 2.3.

The test statistic for a One-Sample *t*-test is:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

$\bar{x}$ = the sample average
s = the standard deviation of the sample values
n = the number of values in the sample
$\frac{s}{\sqrt{n}} = S_{\bar{x}} =$ standard error of the mean

Continuing with the fitness club example, the sample mean calculated from the data above is 2.47 and the sample standard deviation is 1.33. Plugging into the t-statistic formula above, we can calculate a t-statistic value as follows:

$$t = \frac{(2.47-2.3)}{\frac{1.33}{\sqrt{30}}} = 0.69$$

Step 4: Find the Critical Value and Compare to Test Statistic Value

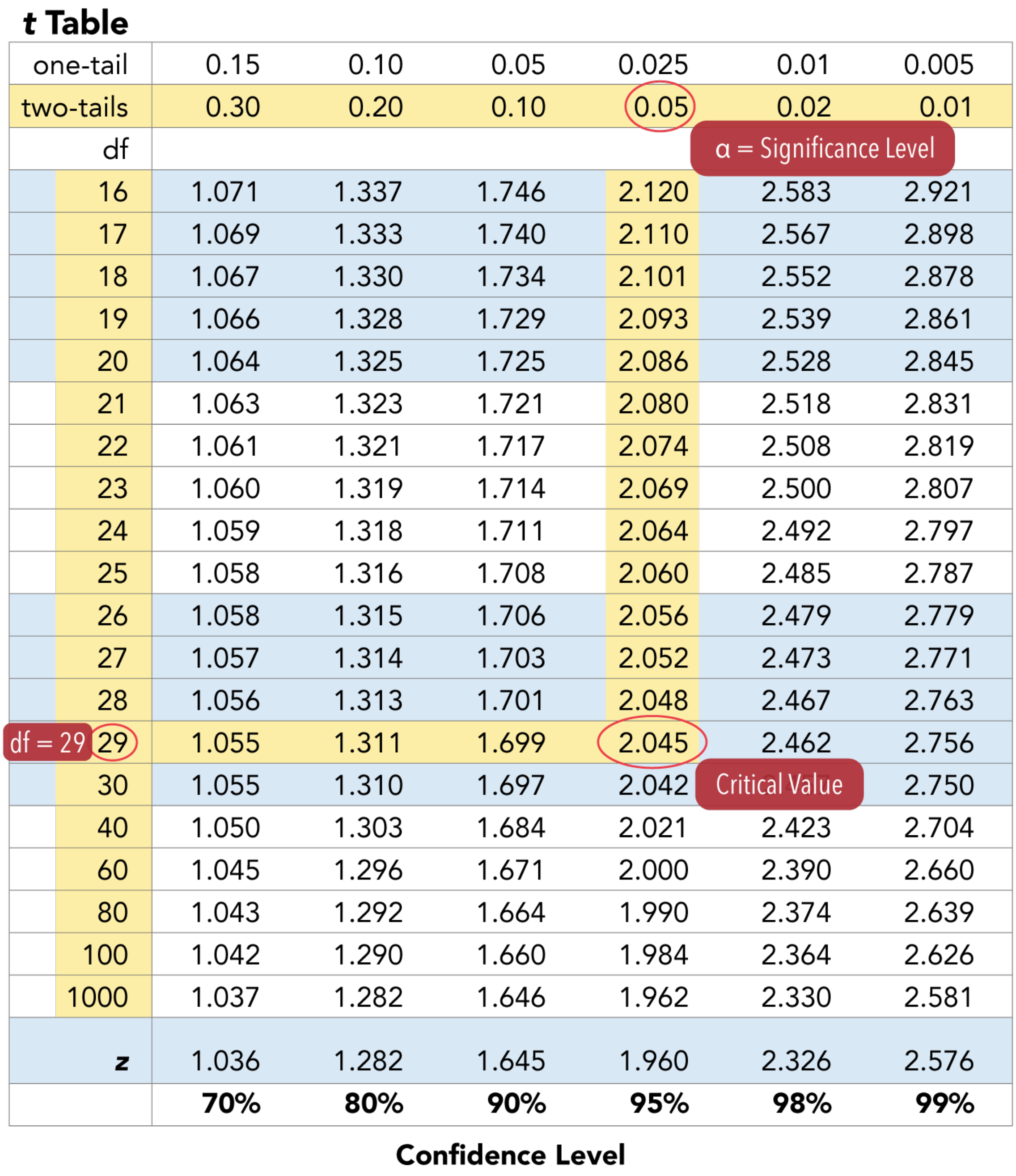
Your critical value will depend upon the test statistic that you employed. We are using a t-test for the fitness club example. For a t-test, the greater the magnitude of your t-test statistic's value, the more likely you are to reject the null hypothesis. It is important to note that, holding n and s constant, the t-test statistic becomes larger as the sample mean becomes larger than μ_0. Likewise, the closer the sample mean is to μ_0, the closer the t-test statistic is to zero.

To determine whether your t-test statistic's value is large enough for you to reject the null hypothesis, you must find the critical values for your hypothesis test (the values at which you decide to reject H_0) from a distribution table. The distribution table you examine corresponds to the test statistic used. The critical value within the table depends upon your degrees of freedom and your significance level.

To find the critical value, you'll need to calculate the degrees of freedom (df) for the test. In this case, the degrees of freedom are found by calculating *n* - 1. In the baseball example, df = 30 - 1 = 29.

We will use the *t*-table given below to determine the critical value for this test. To do this, find the table value in the row corresponding to df=29 and the column for α = 0.05, where α is your chosen significance level.

It is important to note that this is a two-sided test, so we need to look at the column whereα = 0.05 in a two-sided test. This is shown below.



t Table

one-tail	0.15	0.10	0.05	0.025	0.01	0.005
two-tails	0.30	0.20	0.10	0.05	0.02	0.01
df						
16	1.071	1.337	1.746	2.120	2.583	2.921
17	1.069	1.333	1.740	2.110	2.567	2.898
18	1.067	1.330	1.734	2.101	2.552	2.878
19	1.066	1.328	1.729	2.093	2.539	2.861
20	1.064	1.325	1.725	2.086	2.528	2.845
21	1.063	1.323	1.721	2.080	2.518	2.831
22	1.061	1.321	1.717	2.074	2.508	2.819
23	1.060	1.319	1.714	2.069	2.500	2.807
24	1.059	1.318	1.711	2.064	2.492	2.797
25	1.058	1.316	1.708	2.060	2.485	2.787
26	1.058	1.315	1.706	2.056	2.479	2.779
27	1.057	1.314	1.703	2.052	2.473	2.771
28	1.056	1.313	1.701	2.048	2.467	2.763
29	1.055	1.311	1.699	2.045	2.462	2.756
30	1.055	1.310	1.697	2.042		2.750
40	1.050	1.303	1.684	2.021	2.423	2.704
60	1.045	1.296	1.671	2.000	2.390	2.660
80	1.043	1.292	1.664	1.990	2.374	2.639
100	1.042	1.290	1.660	1.984	2.364	2.626
1000	1.037	1.282	1.646	1.962	2.330	2.581
z	1.036	1.282	1.645	1.960	2.326	2.576
	70%	**80%**	**90%**	**95%**	**98%**	**99%**

Confidence Level

α = Significance Level

df = 29

Critical Value

From the *t*-table, we find that the critical value for this specific situation is 2.045.

Because your test statistic's value (0.69) is not greater than 2.045, you do not reject your null hypothesis. What this means is that the difference between your sample mean and μ_0 is not significant. There is not enough evidence for you to conclude that your null hypothesis is false.

The result of the hypothesis test is that you do not have sufficient reason to conclude that the true population mean is now greater than 2.3.

3.06.2 Hypothesis Testing Example, Two-Sample *t*-Test

Hypothesis Testing Example, Two-Sample *t*-Test

Example: Baseball Pitch Speed

In baseball, players and fans are often enamored by the velocity of a pitcher's pitches. Some pitches become ineffective when their velocity decreases by a few miles per hour.

Mike Rodriguez pitches for the New York Scorpions. After a bad performance one afternoon, Rodriguez tries to explain his struggles. "My pitches always have a higher velocity during night games," Rodriguez says. "I've pitched better at night than in the afternoon my entire life!"

A rival hitter, Matt Johnson of the Boston Minutemen, publicly disagrees with Rodriguez's explanation. "Rodriguez is deflecting. He throws harder during day games than night games!"

A savvy reporter decides to study Rodriguez and Johnson's claims. A team scout is keeping a log of Rodriguez' pitch velocity throughout the season. There is no information available for prior seasons, so the reporter can only use this season as a sample. In fact, there are two samples. One sample is Rodriguez's afternoon games, while the other sample is Rodriguez's night games:

Rodriguez' Night Pitch Velocity

Date	Average Pitch Velocity (mph)
May 17	89
May 23	90
May 29	90
June 4	95
June 10	94
June 16	94
June 26	91
July 7	93

Rodriguez' Afternoon Pitch Velocity

Date	Average Pitch Velocity (mph)
May 7	86
May 12	88
June 21	92
July 2	88
July 12	86
July 19	90
July 25	87
July 30	86

Hypothesis Testing Steps, Baseball Example

- Step 1
- Step 2
- Step 3
- Step 4

Step 1: State the Null Hypothesis and the Alternative Hypothesis

For our example, the null hypothesis is: "Rodriguez's pitch velocity in afternoon games is equal to his pitch velocity in night games." The null hypothesis states that there is **no significant difference** between these two samples. The null hypothesis can be shown as:

$$H_0 : \mu_1 = \mu_2$$

To test the validity of the null hypothesis, you must specify an alternative hypothesis, H_A.

Your alternative hypothesis is: "Rodriguez's pitch velocity in afternoon games is less than his pitch velocity in night games." Note that the player is claiming his pitches are faster at night, so we can specify that claim in the alternative hypothesis. The alternative hypothesis states that there **is a significant difference** between these two samples. The alternative hypothesis can be written as:

$$H_A : \mu_1 < \mu_2$$

This is an appropriate situation to use a one-tailed test because the claim is that Rodriguez pitches *faster* during night games.

Step 2: Decide on the Significance Level

The significance level is a decision criterion that specifies the degree of certainty with which you want to make your judgment of whether or not to reject the null hypothesis. The significance level is the probability that you will mistakenly reject a true null hypothesis based on the sample statistic. The more careful you want to be about not rejecting a true null hypothesis, the smaller your significance level should be:

- A **higher significance level** indicates a **higher threshold to reject the null hypothesis**. To state that there is a significant difference, you have to be more certain that random chance or error is not causing the difference.

- A **lower significance level** indicates a **lower threshold to reject the null hypothesis**. To state that there is a significant difference, you do not have to be as certain that random chance or error is not causing the difference.

A commonly used significance level in many research settings is 0.05. A 0.05 significance level means that you state that the results were significant if there is only a 5% chance it was actually caused by random variation or errors. You expect a rejected null hypothesis to be an incorrect decision in only 5% of cases. We will use a 0.05 significance level in this example.

Step 3: Compute the Value of the Test Statistic

Be sure to choose an appropriate test statistic for your hypothesis test.

For a hypothesis about two sample means, hypothesis testing allows you to determine if there is a significant difference between the two sample means. For the baseball example, the further the two sample means are from one another (in addition to other factors such as sample size and variance), the more reason there is to reject the null hypothesis.

To measure whether the difference between the two sample means is significant, we'll need to employ a test statistic. The most useful test statistic for this scenario is the Two-Sample *t*-Test. The Two-Sample *t*-Test is used to compare two sample means to one another.

The test statistic for a Two-Sample *t*-Test is:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{Sp\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where:

$\bar{x}_1$ = the sample mean for Sample 1
$\bar{x}_2$ = the sample mean for Sample 2

S_p = the pooled standard deviations of both samples = $\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

n_1 = the number of data points in Sample 1
n_2 = the number of data points in Sample 2

Continuing with the baseball example, we can calculate and plug in values from the data above. Here, x_1 is the night game sample, while x_2 is the afternoon game sample.

From the data above, we can calculate the sample mean and standard deviation for each sample. The sample mean and standard deviation for the afternoon games are 87.875 and 2.167, respectively. The sample mean and standard deviation for the evening games are 92 and 2.268, respectively.

Each sample has a size of 8, son_1 and n_2 are both 8.

We first calculate the pooled standard deviation, s_p:

$$S_p = \sqrt{\frac{(8-1)(2.167)^2 + (8-1)(2.268)^2}{8+8-2}} = 2.218$$

Plugging these values into the *t*-statistic formula above, we can calculate a *t*-statistic value as follows:

$$t = \frac{92-87.875}{2.218\sqrt{\frac{1}{8}+\frac{1}{8}}} = 3.72$$

Step 4: Find the Critical Value and Compare to Test Statistic Value

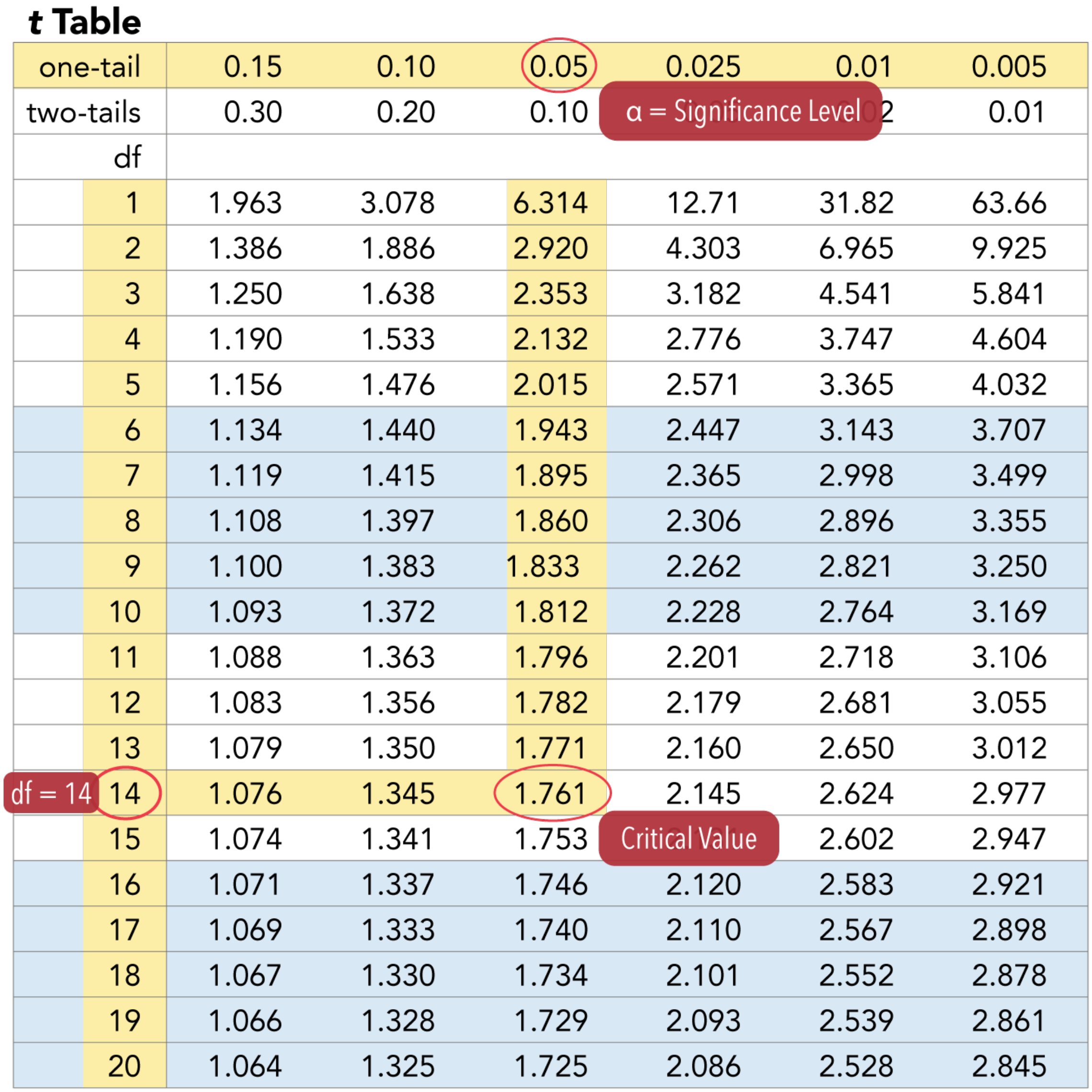
Your critical value will depend upon the test statistic that you employed. We are using a *t*-Test for the baseball example. For a *t*-Test, the greater the magnitude of your *t*-Test statistic's value, the more likely you are to reject the null hypothesis. It is important to note that holding all else equal, the *t*-Test statistic becomes larger as the sample means, $\bar{x}_1$ and $\bar{x}_2$, become further apart. Likewise, the closer the sample means are to one another, the closer the *t*-Test statistic is to zero.

To determine whether your *t*-Test statistic's value is large enough for you to reject the null hypothesis, you must find the critical values for your hypothesis test (the values at which you decide to reject H_0) from a distribution table. The distribution table you examine corresponds to the test statistic used. The critical value within the table depends upon your degrees of freedom and your significance level.

To find the critical value, you'll need to calculate the degrees of freedom (df) for the test. In this case, the degrees of freedom are found by calculating $n_1 + n_2 - 2$. In the baseball example, df = 8 + 8 - 2 = 14.

We will use the *t*-table given below to determine the critical value for this test. To do this, find the table value in the row corresponding to df=14 and the column for α = 0.05, where α is your chosen significance level.

It is important to note that this is a one-sided test, so we need to look at the column where α = 0.05 in a one-sided test. This is shown below.



t Table

one-tail	0.15	0.10	0.05	0.025	0.01	0.005
two-tails	0.30	0.20	0.10	[illegible]	0.02	0.01
df						
1	1.963	3.078	6.314	12.71	31.82	63.66
2	1.386	1.886	2.920	4.303	6.965	9.925
3	1.250	1.638	2.353	3.182	4.541	5.841
4	1.190	1.533	2.132	2.776	3.747	4.604
5	1.156	1.476	2.015	2.571	3.365	4.032
6	1.134	1.440	1.943	2.447	3.143	3.707
7	1.119	1.415	1.895	2.365	2.998	3.499
8	1.108	1.397	1.860	2.306	2.896	3.355
9	1.100	1.383	1.833	2.262	2.821	3.250
10	1.093	1.372	1.812	2.228	2.764	3.169
11	1.088	1.363	1.796	2.201	2.718	3.106
12	1.083	1.356	1.782	2.179	2.681	3.055
13	1.079	1.350	1.771	2.160	2.650	3.012
14	1.076	1.345	1.761	2.145	2.624	2.977
15	1.074	1.341	1.753	[illegible]	2.602	2.947
16	1.071	1.337	1.746	2.120	2.583	2.921
17	1.069	1.333	1.740	2.110	2.567	2.898
18	1.067	1.330	1.734	2.101	2.552	2.878
19	1.066	1.328	1.729	2.093	2.539	2.861
20	1.064	1.325	1.725	2.086	2.528	2.845

α = Significance Level

df = 14

Critical Value

From the *t*-table, we find that the critical value for this specific situation is 1.761.

Because your test statistic's value (3.72) is greater than 1.761, you reject your null hypothesis. What this means is that the difference between your two sample means **is significant**.

The result of the hypothesis test is that you do have sufficient reason to conclude that the two means are different.**Mike Rodriguez does throw faster at night than during the day!** Statistically, his claim is correct.

3.06.3 Managerial Statistics Scenario

This assignment does not contain any printable content.

3.07 Chi-Squared Test

Chi-Squared Test

A chi-squared test (also written as "X^2" or " **chi-square**") is a common hypothesis test. Like the one-sample t-test, or the two-sample t test, a chi-squared test is commonly used in statistics to draw inferences about a population, by testing sample data.

Interestingly, a chi-squared test is employed for categorical data. Categorical data breaks results into categories, like days of the week, or states of the United States of America.

You may recall that any hypothesis test has both a null hypothesis and an alternative hypothesis. Chi-square tests are no different. The null hypothesis, or H_0, is the statement that there is no relationship. The alternative hypothesis, or H_A, is the opposite statement to the null hypothesis. It states that there is a relationship for whatever relationship is being tested.

When performing a chi-square test, the **null hypothesis** is the statement that there is no significant difference between the distribution of your data and a specified distribution. Put simply, the null hypothesis states that *the data is distributed as expected.*

The **alternative hypothesis** in a chi-squared test is the statement that there is a significant difference between the distribution of your data and a specified distribution. Put simply, the alternative hypothesis states that *the data is **not** distributed as expected.*

All hypothesis tests follow the same four steps:

- **Step 1:** State the Null Hypothesis and the Alternative Hypothesis
- **Step 2:** Decide on the Significance Level
- **Step 3:** Compute the Value of the Test Statistic
- **Step 4:** Find the Critical Value and Compare to Test Statistic Value

To see, how a chi-square test is used, let's dive into each of these steps, as they pertain to a chi-square case study about ice cream!

Ice Cream Case Study

Lisa has recently been placed in charge of buying chocolate, vanilla, and strawberry ice cream for a weekly ice cream party in the office. Previously, Will would purchase an equal amount of vanilla and strawberry ice cream, but twice as much chocolate ice cream as vanilla or strawberry ice cream. Put another way, of the ice cream Will purchased, 50% would be chocolate, 25% vanilla, and 25% strawberry.

Being a perfectionist, Lisa wanted to make sure she was purchasing the right distribution of each flavor of ice cream. She doesn't mind purchasing too much ice cream: there can always be left-overs. What Lisa wants to avoid, though, is purchasing the wrong proportion of each flavor. So, she needs to ensure that her purchases are proportional to her co-workers' demand for each flavor.

First, Lisa decides to perform a chi-squared test on this week's ice cream party. She needs to specify her null and alternative hypotheses, to understand what is being tested.

Chi Squared Hypothesis Testing Steps, Ice Cream Case Study

Step 1: State the Null Hypothesis and the Alternative Hypothesis

- The null hypothesis, or H_0, is the statement that the data is distributed as expected.
- The alternative hypothesis, or H_A, is the statement that the data is **not** distributed as expected.

The null hypothesis, or H_0, is the statement that the data is distributed as expected. In this case, that null hypothesis is: the demand for ice cream flavors follows the same distribution as Will's purchases — 50% chocolate, 25% vanilla, and 25% strawberry.

Lisa's alternative hypothesis, or H_A, is the statement that the data is **not** distributed as expected. Here, that alternative hypothesis is: the demand for ice cream flavors does **not** follow the same distribution as Will's purchases.

Step 2: Decide on the Significance Level

The significance level is a decision criterion that specifies the degree of certainty with which you want to make your judgment of whether or not to reject the null hypothesis. The significance level is the probability that you will mistakenly reject a true null hypothesis based on the sample statistic. The more careful you want to be about not rejecting a true null hypothesis, the smaller your significance level should be:

Lisa decides to use a common significance level: 0.05. A 0.05 significance level means that she states that the results were significant if there is only a 5% chance it was actually caused by random variation or errors. Lisa expects a rejected null hypothesis to be an incorrect decision in only 5% of cases.

Step 3: Compute the Value of the Test Statistic

Chi-Squared Test Formula

The formula for the chi-squared test is:

$$X^2 = \Sigma \frac{(O-E)^2}{E}$$

where:

O = the observed value in any given category.
E = the expected value in any given category.

So, $O - E$ is the difference between observed and expected values in any category.

X^2 is the sum of $\frac{(O-E)^2}{E}$, for all categories.

Determining Observed Values

Lisa knows that $X^2 = \Sigma \frac{(O-E)^2}{E}$, for each of the ice cream flavors.

To compute the X^2 value, Lisa first needs to determine the observed (O) and expected (E) values for each of the ice cream flavors, individually.

For the observed value for each flavor, Lisa observes the following during the party:

Ice Cream Flavor	Number of Scoops Ordered
Chocolate	54
Vanilla	29
Strawberry	41
Total	124

Determining Expected Values

Now, she needs to determine the expected values (E) for each flavor. Lisa expected 50% of the scoops to be chocolate, 25% of the scoops to be vanilla, and 25% of the scoops to be strawberry. From the total number of scoops ordered (124) Lisa is able to calculate the expected values of chocolate, vanilla, and strawberry (E_c, E_v, and E_s, respectively):

Chocolate: $E_c = 124 \times 0.5 = 62$
Vanilla: $E_v = 124 \times 0.25 = 31$
Strawberry: $E_s = 124 \times 0.25 = 31$

Calculating X^2

Next, Lisa plugs in these values to calculate $\frac{(O-E)^2}{E}$ for each of the ice cream flavors:

- **Chocolate:** $\frac{(O_c-E_c)^2}{E_c}$
  $\frac{(54-62)^2}{62}$
  $= 1.0323$
- **Vanilla:** $\frac{(O_v-E_v)^2}{E_v}$
  $\frac{(29-31)^2}{31}$
  $= 0.1290$
- **Strawberry:** $\frac{(O_s-E_s)^2}{E_s}$
  $\frac{(41-31)^2}{31}$
  $= 3.2258$

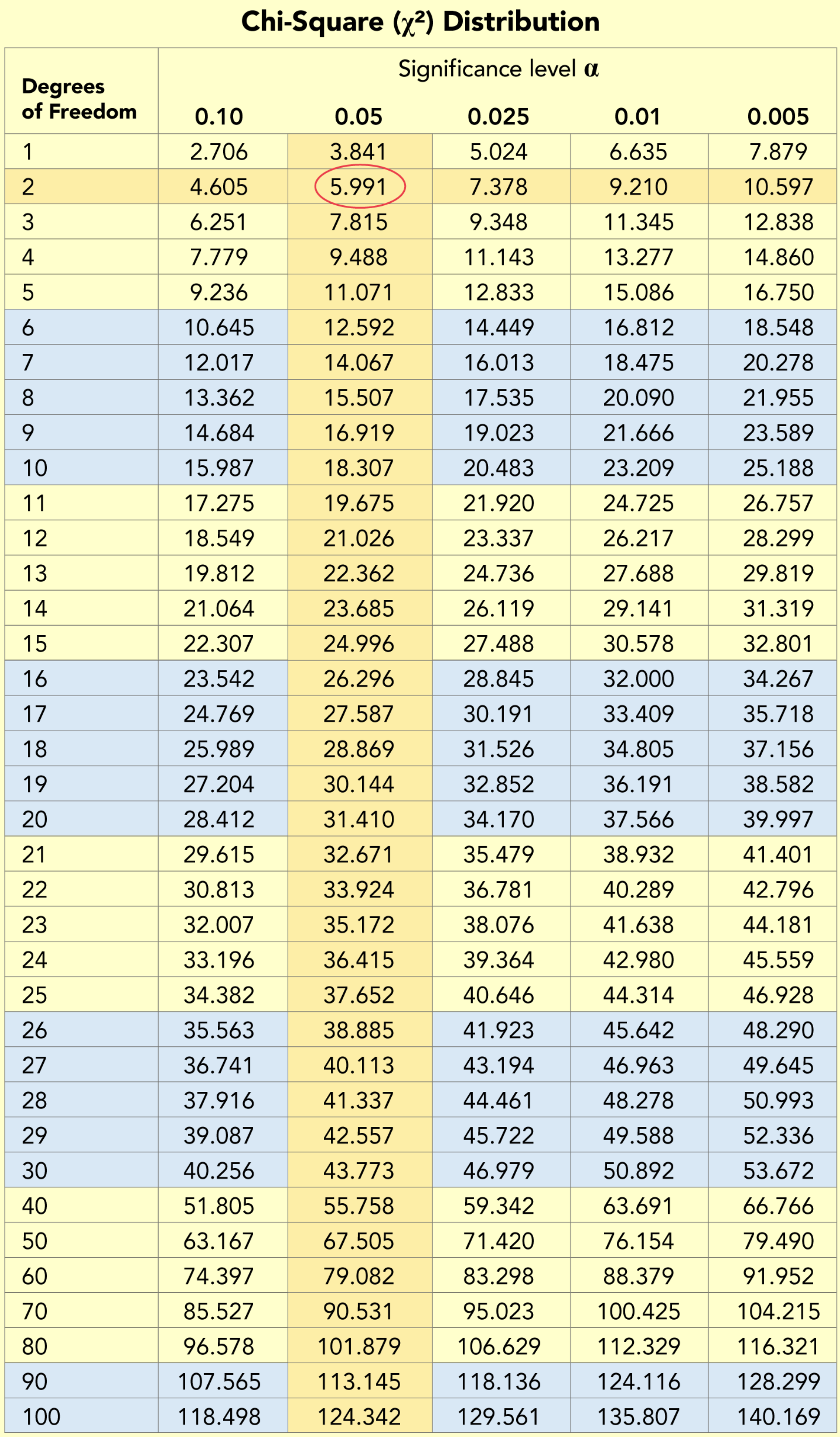
The X^2 value is the sum of these three values.$X^2 = 1.0323 + 0.1290 + 3.2258 =$ **4.387**

Step 4: Find the Critical Value and Compare to Test Statistic Value

As with other hypothesis testing, to determine whether your test statistic's value is large enough for you to reject the null hypothesis, you must find the critical values for your hypothesis test (the values at which you decide to reject H_0) from a distribution table. The distribution table you examine corresponds to the test statistic used, which is Chi-Square here.

To find the critical value, you'll need to calculate the degrees of freedom (df), which is the sample size minus one(n - 1). For a chi-squared test, the sample size is the number of categories. So, for this ice cream example, the degrees of freedom = (3 - 1) = 2.

Lisa determines her Chi-Squared test's critical value by finding the table value in the row corresponding to df= 2 and the column for α = 0.050, where α is her chosen significance level. From the Chi-Square table, she find that her critical value is 5.99146.



Chi-Square (χ^2) Distribution

Degrees of Freedom	Significance level α				
	0.10	**0.05**	**0.025**	**0.01**	**0.005**
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.071	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188
11	17.275	19.675	21.920	24.725	26.757
12	18.549	21.026	23.337	26.217	28.299
13	19.812	22.362	24.736	27.688	29.819
14	21.064	23.685	26.119	29.141	31.319
15	22.307	24.996	27.488	30.578	32.801
16	23.542	26.296	28.845	32.000	34.267
17	24.769	27.587	30.191	33.409	35.718
18	25.989	28.869	31.526	34.805	37.156
19	27.204	30.144	32.852	36.191	38.582
20	28.412	31.410	34.170	37.566	39.997
21	29.615	32.671	35.479	38.932	41.401
22	30.813	33.924	36.781	40.289	42.796
23	32.007	35.172	38.076	41.638	44.181
24	33.196	36.415	39.364	42.980	45.559
25	34.382	37.652	40.646	44.314	46.928
26	35.563	38.885	41.923	45.642	48.290
27	36.741	40.113	43.194	46.963	49.645
28	37.916	41.337	44.461	48.278	50.993
29	39.087	42.557	45.722	49.588	52.336
30	40.256	43.773	46.979	50.892	53.672
40	51.805	55.758	59.342	63.691	66.766
50	63.167	67.505	71.420	76.154	79.490
60	74.397	79.082	83.298	88.379	91.952
70	85.527	90.531	95.023	100.425	104.215
80	96.578	101.879	106.629	112.329	116.321
90	107.565	113.145	118.136	124.116	128.299
100	118.498	124.342	129.561	135.807	140.169

Because her test statistic's value (4.387) is not greater than 5.99146, she does not reject the null hypothesis. This means that she can not conclude that the ice cream demand distribution is different from the purchasing distribution.

Based on these results, Lisa decides to continue purchasing 50% chocolate ice cream, 25% vanilla ice cream, and 25% strawberry ice cream.

3.08 ANOVA

ANOVA

Analysis of Variance (ANOVA) is a technique used to determine if there is a significant difference among three or more means. Using ANOVA, we see if there is sufficient evidence from sample data of three or more populations to determine whether the population means are all equal, or whether there is a significant difference with at least one of the means. Ordinary hypothesis testing can be used to determine whether there is a basis to conclude that two populations have different means, but ANOVA is helpful in determining a significant difference among three or more means.

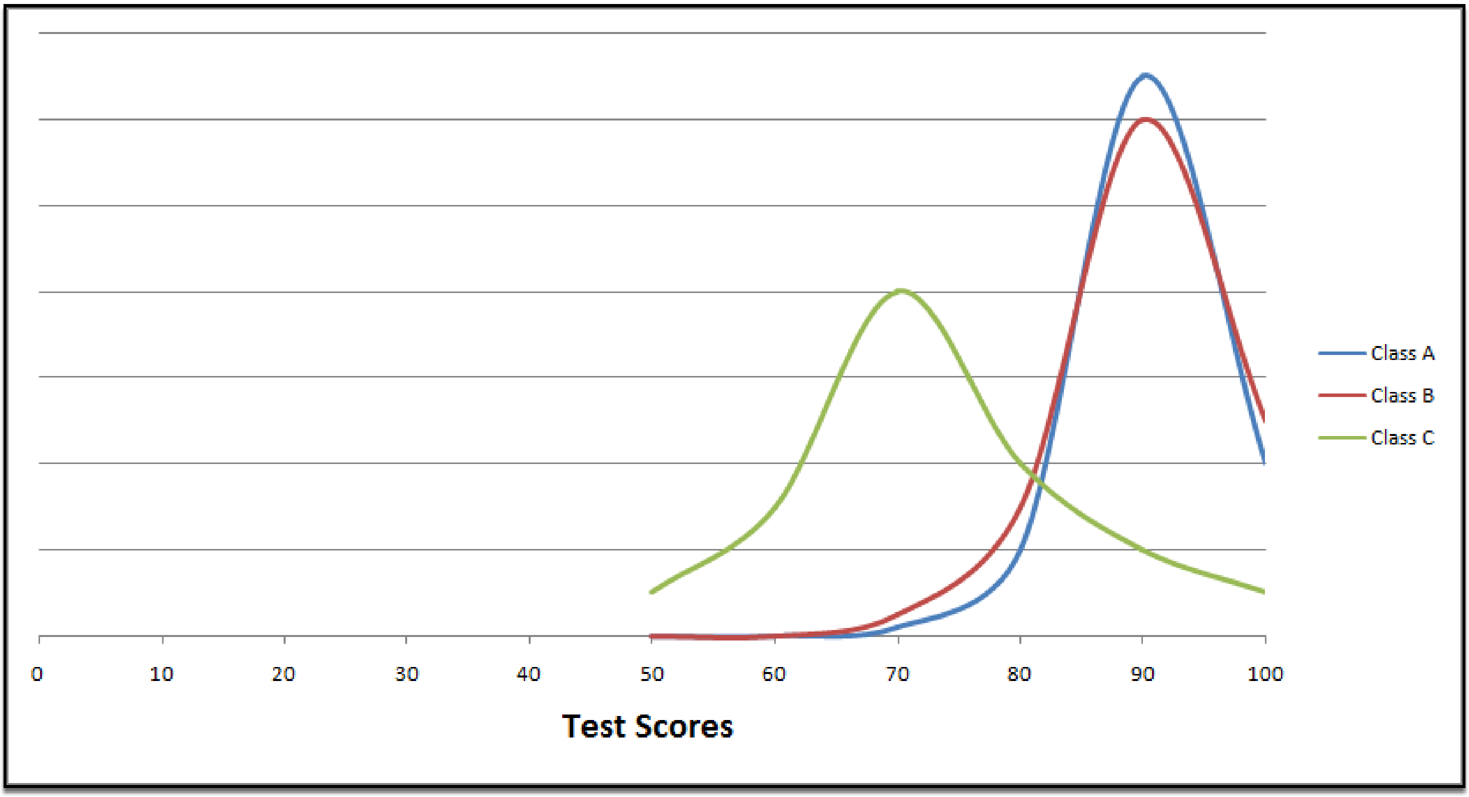
The null hypothesis claims that all population means are equal. For example, if three populations are being tested, the null hypothesis would be H_0: $\mu_1 = \mu_2 = \mu_3$.

The alternative hypothesis states that not all of the population means are equal. We accept the alternative hypothesis if at least one of the population means is considered significantly different.

An F-value is the test statistic that is utilized in ANOVA. As we've seen with other test statistics, our test statistic value and the critical value determine whether we "reject the null hypothesis" or "fail to reject the null hypothesis." As always, if our test statistic value exceeds the critical value, we reject the null hypothesis. We would conclude that at least one of the population means is significantly different from the others.

Test Scores

The following graph shows test result distributions from three different classes. ANOVA is a useful tool for testing the difference among multiple sets of data. Examine the graph below and decide which class' results are different or similar to one another:



We can see that the datasets from Class A and Class B are very similar to one another. On the other hand, the distribution of results in Class C are different from Class A, or Class B. Class C clearly has a greater variance and a lower mean. ANOVA is a widely used to tool to quantify the difference in means among data sets.

Imagine we only have a random sample of test scores from each class. From this information, we want to use ANOVA to analyze the differences and similarities among the groups. Below, we have a random sample of students from each class.

Sample Test Score by Class

Class A	Class B	Class C
81	83	60
87	88	68
91	91	74
94	94	82
100	99	94

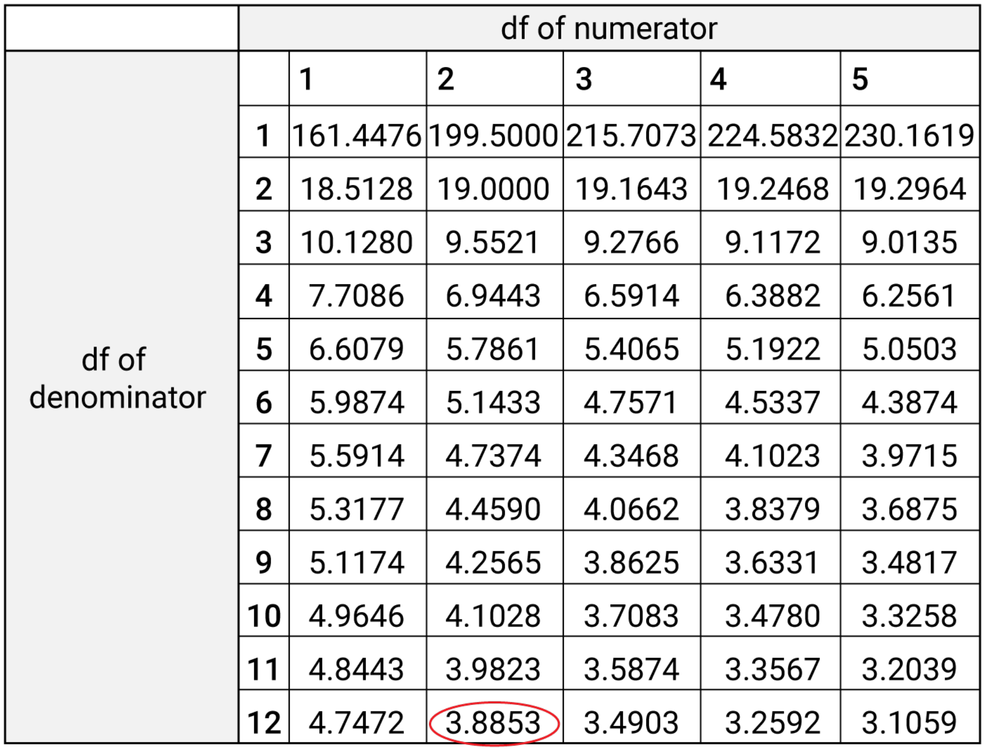
SUMMARY

Group	Count	Sum	Average	Variance
Class A	5	453	91	51.3
Class B	5	455	91	36.5
Class C	5	378	76	170.8

ANOVA

Source of Variance	degrees of [illegible] th [illegible] of [illegible] is [illegible] er [illegible] n [illegible] e, [illegible] the [illegible] l is [illegible] o [illegible] n [illegible] sa [illegible] data sets.	Sum o [illegible] S [illegible] es [illegible] th [illegible] of [illegible] t [illegible] l [illegible] the [illegible] of best fit.	[illegible] re [illegible] n [illegible] o [illegible] s [illegible] re [illegible] d [illegible] th [illegible] he [illegible] me [illegible] he [illegible] re [illegible] g [illegible] of the residual.	F [illegible] e [illegible] m [illegible] on [illegible] and [illegible] the [illegible] va [illegible] os [illegible] gre [illegible] een [illegible] by using [illegible] istribution table.	va [illegible] of [illegible] re [illegible] he [illegible] p [illegible] a [illegible] indic [illegible] cally [illegible] st [illegible] es, [illegible] in [illegible] es [illegible] a [illegible] ly [illegible] re [illegible] e [illegible] variable.
Between G [illegible] n analy [illegible], this repr [illegible] ents [illegible] betw [illegible] the datasets.	2	770.5	385.3	4.469	0.03543
Within G [illegible] n analy [illegible], this repr [illegible] ents [illegible] with [illegible] ll of datasets.	12	1034.4	86.2		
Total	19	30			

For these three samples, the test-statistic has a value of F = 4.469. Referring to the F-distribution table, using a five percent level of significance (*α* = 0.05), a degree of freedom for the numerator of 2 (number of groups minus one) and a degree of freedom of 12 for the denominator (total number of observations minus the number of groups) we find that the critical F-value of 3.89.



		df of numerator				
		1	2	3	4	5
df of denominator	1	161.4476	199.5000	215.7073	224.5832	230.1619
	2	18.5128	19.0000	19.1643	19.2468	19.2964
	3	10.1280	9.5521	9.2766	9.1172	9.0135
	4	7.7086	6.9443	6.5914	6.3882	6.2561
	5	6.6079	5.7861	5.4065	5.1922	5.0503
	6	5.9874	5.1433	4.7571	4.5337	4.3874
	7	5.5914	4.7374	4.3468	4.1023	3.9715
	8	5.3177	4.4590	4.0662	3.8379	3.6875
	9	5.1174	4.2565	3.8625	3.6331	3.4817
	10	4.9646	4.1028	3.7083	3.4780	3.3258
	11	4.8443	3.9823	3.5874	3.3567	3.2039
	12	4.7472	3.8853	3.4903	3.2592	3.1059

The rejection region (the area in which we reject the null hypothesis) is therefore*F* > 3.89. Our test-statistic is 4.469, so we find the test-statistic does lie in the rejection region. Therefore, **we reject the null hypothesis.** This means that at least one of the data sets were found to be significantly different from the others. The sample provides sufficient evidence to conclude that **at least one population mean is significantly different from the others.**

As you can see above, Class A and Class B are very similar, while Class C is different. It is quite likely that Class C is the cause for the significant difference. When performing ANOVA, the null hypothesis is rejected if *at least one* sample is significantly different. So, even though Class A and Class B may not have a significant difference, the fact that Class C is different is enough for us to reject the null hypothesis.

3.09 Video: Different Statistical Techniques

This assignment does not contain any printable content.

3.10 Forecasting, Regression Analysis, and Quantitative Techniques

Forecasting, Regression Analysis, and Quantitative Techniques

Managers use different quantitative techniques to evaluate data, see patterns and trends and make decisions involving pricing, cost, and market viability.

The following chart provides a summary of several common techniques, which will be covered in greater detail in this module.

Forecasting Techniques

Forecasting and Regression analysis	Forecasting is one of the most important elements of business decision making. Managers employ three basic forecasting techniques: Judgmental (based on sales, consumer, or management input); Time-Series (based upon data patterns in past data, which includes techniques for random variation, trend, seasonality, etc.) and Associative (based upon predictive or explanatory variables and includes regression.) We're focusing on regression analysis because it is often cited as a powerful technique and managers need to understand it. *Regression Analysis definition:* statistical method to measure the average amount of change in a dependent variable associated with a unit change in one or more independent variables; considered an associate model as it incorporates the factors (variables) that might influence the quantity being forecasted *Used for:* analyzing cost behavior and forecasting future sales
Other Quantitative Techniques	
Time series analysis	*Definition:* forecasting technique that employs a series of past data points to make a forecast *Used for:* evaluating patterns in data to make decisions about staffing levels, inventory, etc.
Cluster analysis	*Definition:* the process of arranging terms or values based on different variables into "natural" groups *Used for:*understanding the makeup of an industry's different areas
Decision analysis	*Definition:* the process of weighing all outcomes of a decision to determine the best course of action *Used for:* making decisions, whether personal or professional

3.11 Benefits and Shortcomings of Regression Analysis and Quantitative Techniques

Benefits and Shortcomings of Regression Analysis and Quantitative Techniques

The following chart summarizes some of the benefits and shortcomings of these techniques.

	Benefits	**Shortcomings**
Regression analysis	- Allows sophisticated analysis of cost behavior and sales forecasts - Provides objective benchmarks for evaluation of reliability of estimates	- Requires 15 or more data points for accuracy - Can be influenced by outliers (unusual data points) - Requires informed analysis
Time series analysis	- Aids decision making by finding patterns in data, such as sales trends - Allows performance and productivity evaluation	- Assumes past data patterns will repeat in future, which may not be true - Key variables may not be captured
Cluster analysis	- Sorts individual data points into different groups - Helps determine target markets - Identifies successful and unsuccessful habits and systems	- Long and expensive process - There are hundreds of potential approaches to take, each specific to a certain situation
Decision analysis	- Determines the decision with the greatest value - Produces a value under certainty, uncertainty, and risk	- Quality of decision is limited to the amount of data available - Does not emphasize the risk of the worst case scenario

3.12 Video: Regression Analysis

This assignment does not contain any printable content.

3.13 Linear Regression

Regression Analysis

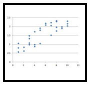
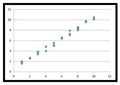
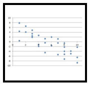
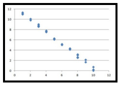
When carrying out a forecast analysis, we find that there are usually multiple variables that interact with each other. Regression analysis is used when multiple variables' quantities relate to each other. For example, a home sale price may be related to its square footage or its walking distance to a school or train station. Here, sale price, square footage, distance from a school and distance from a train station are all variables that can be modeled with a regression analysis. Once their relationship has been established, regression analysis can be used to predict future data values.

Dependent vs. Independent Variables

The dependent variable is the variable whose value depends on the other variables in the equation; typically the cost or activity to be predicted (in the previous example, the dependent variable was home sale price). The independent variables are variables presumed to influence the dependent variable.

Linear Regression

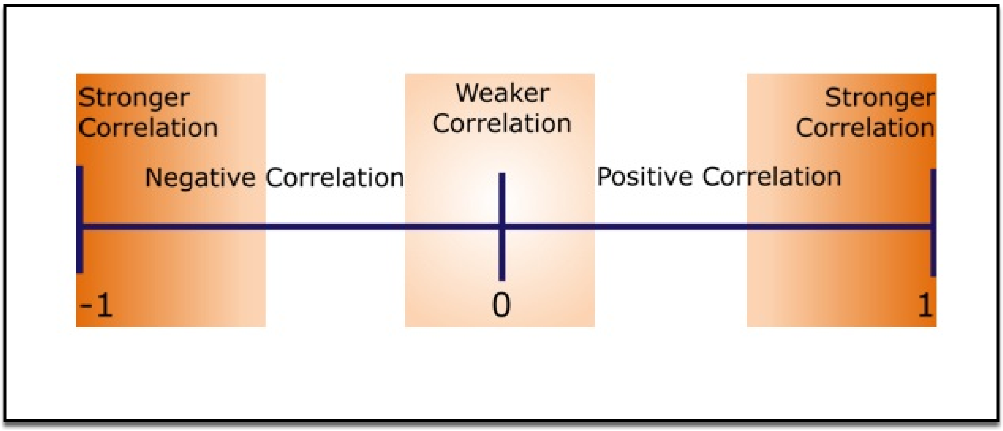
A linear relationship between two variables can be measured by its strength. A strong linear relationship indicates that the data will bunch around a straight line, while a weak linear relationship does not. Although it will rarely follow a straight line exactly, if a regression does fall precisely along a straight line, we call it a perfect linear relationship. If a linear relationship exists, it can be either positive or negative. A positive linear relationship exists when both variables increase together. A negative linear relationship exists when the one variable decreases as the other variable increases. Here some examples of scatterplots with weak, strong, negative and positive linear relationships:

	WEAK	**STRONG**
Positive		
Negative		

The stronger linear relationships hold closer to a line while the weaker relationships do not. Also, the negative relationships move down and to the right, while the positive relationships move up and to the right. Scatterplots are useful for viewing whether a relationship between variables loosely exists.

Correlation

The strength of a linear relationship can be measured with the correlation coefficient. A correlation coefficient, a number between -1 and 1, is only useful in measuring linear regression, rather than nonlinear regression. A correlation coefficient that is close to 0 indicates a weak linear relationship, while a correlation coefficient closer to -1 or 1 represents a strong linear relationship. A correlation coefficient equal to exactly -1 or 1 would be considered perfectly linear. In the case of a perfectly linear relationship, our data points would form a perfectly straight line. Negative linear relationships have correlation coefficients less than 0. Positive linear relationships have correlation coefficients greater than 0.



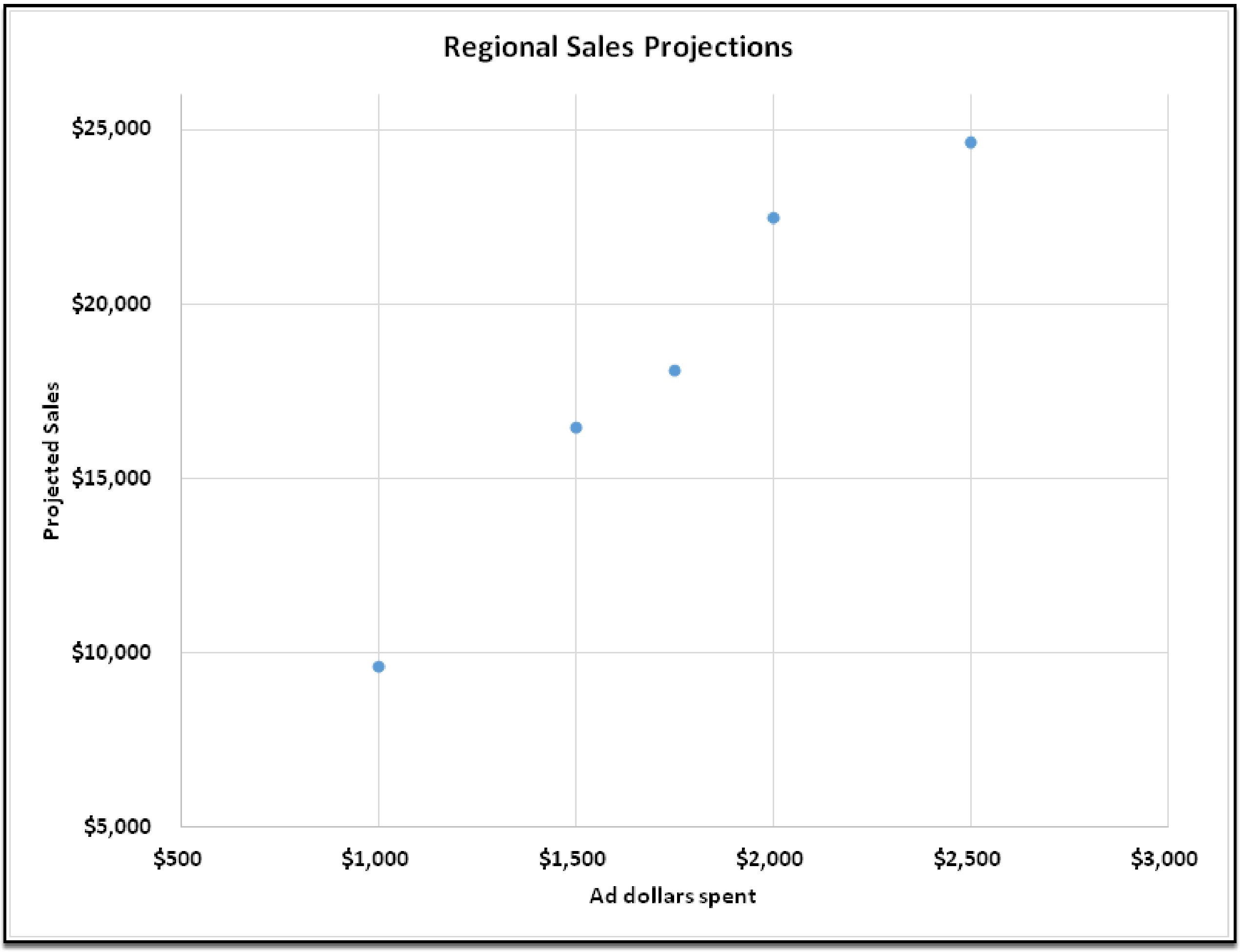
An Example of Regression Analysis

To predict future sales figures based on the amount of money spent on regional advertising, we collect the following data:

Ad Dollars Spent (x)	Sales (y)
$1,000	$9,655
$1,500	$16,500
$1,750	$18,110
$2,000	$22,485
$2,500	$24,670

If plotted on a graph, these points wouldn't form a line, but instead form a "scatter plot." Any regression analysis begins by creating a scatter diagram or scatterplot.

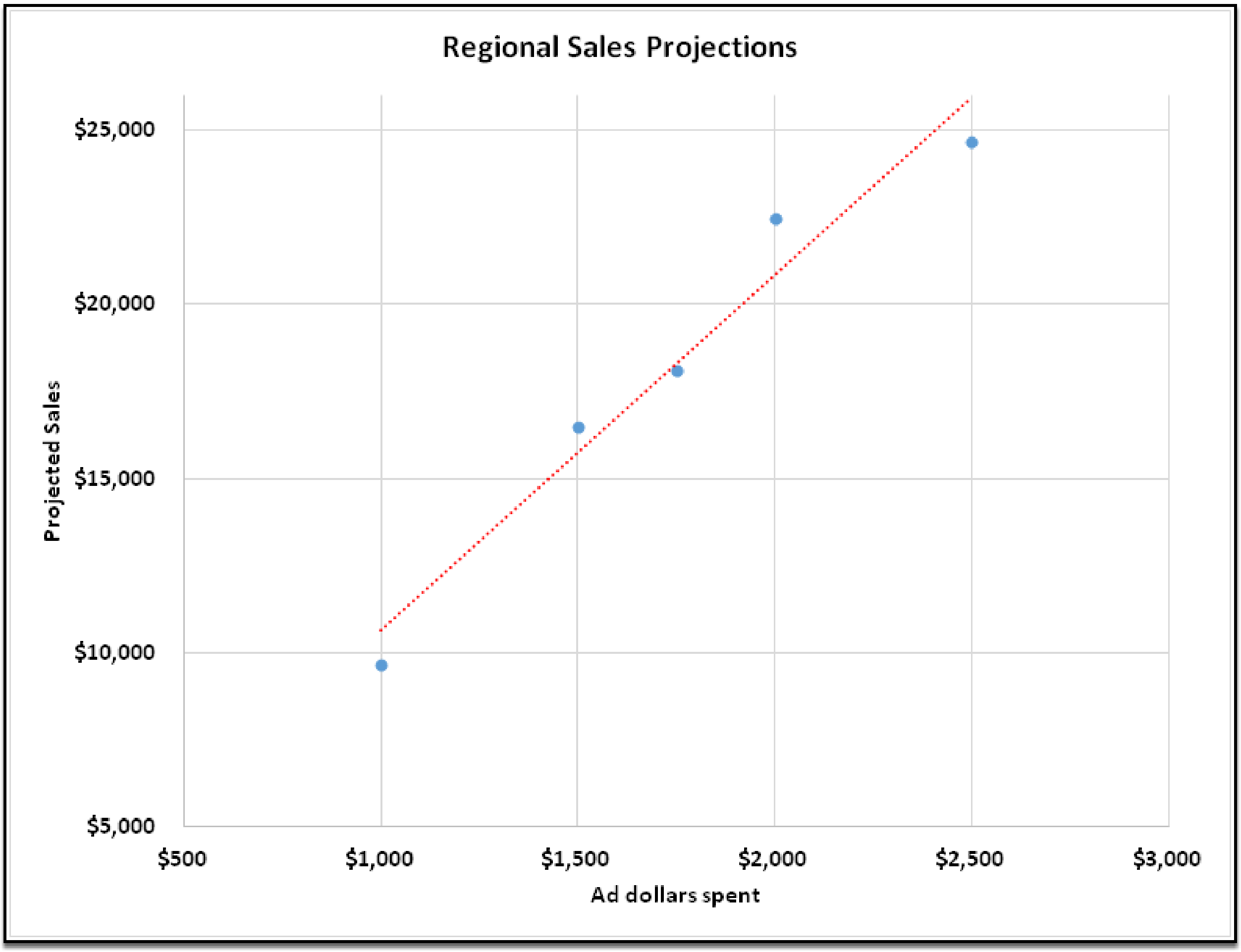
We place the independent variable on the horizontal x-axis. On the vertical y-axis, we place the variable we are testing, the dependent variable. The dependent, response variable responds to the changes of the independent, explanatory variable.



In this example, there is not a perfect "functional" relationship between sales and advertising, where each level of advertising would yield one unique level of sales. However, there does appear to be a statistical relationship between these variables. Each level of sales is associated with a range of possible sales values.

To represent the statistical relationship, we can impose a line through the middle of the points on the scatter plot. This line is called the "line of best fit" or regression line for the data and represents a function. The regression line is the function that best minimizes the distance between each data point and the line, hence the term "line of best fit."

On this graph, the dotted red line is the regression line, and the blue points are our actual data points.

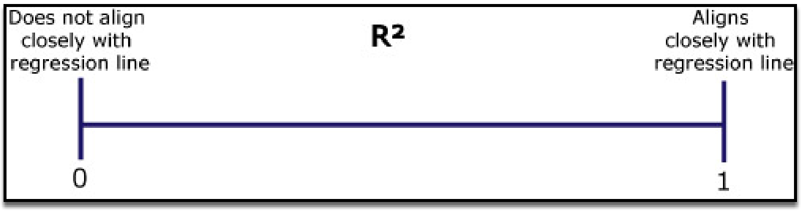


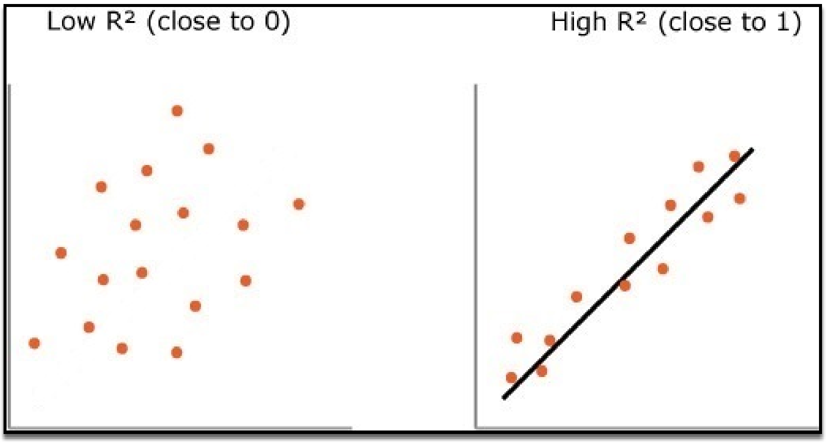
Once the regression line is established, it can be used as a model to predict future outcomes. For example, when modeling a product's sales based on its price, the regression line can be used to determine future sales and to set a price for a product that will make the company the most profit.

The R^2 Statistic

In statistics, the term "R-squared" or "R^2" provides a measure of "goodness of fit."R^2 ranges in value from 0 to 1. An R-squared value close to 1 indicates that the estimation error is small and our data closely aligns to the regression line. If there is only one independent variable and one dependent variable, r-squared is the correlation coefficient squared (*r* = correlation coefficient).

If R-squared is close to 0, our data does not align as closely to the estimated regression line. When a regression line aligns closely to our data (R-Square is high), we can use the equation for that line to make predictions with a high level of confidence.


Does not align closely with regression line
R²
Aligns closely with regression line
0
1



The graph on the right has an R^2 value that is close to 1. Therefore, the line has a high "goodness-of-fit." You can see that the data aligns very closely to the regression line.

The "Low R^2" graph on the left provides a stark contrast to the "High R^2" graph. You can hardly discern any correlation at all! When the data does not align closely with a particular regression line, as is the case in the left-most graph, that indicates that the R^2 value is closer to 0.

Standard Error (SE) of Estimate

In multiple linear regression, there are ways to measure how well the least squares line fits our data. One important way of measuring the fit of our estimated regression line to our dataset is a process called the standard error of estimate. Standard error (SE) of estimate, denoted s^e, is the average deviation of the data points from the predictive regression line or curve. A measure of the accuracy of a prediction driven by a regression model is the standard error of estimate. This differs from the standard error of measurement which measures the amount of variation in the **actual** data around the fitted regression function. Similar to a standard deviation, we know that about $\frac{2}{3}$ of the predictions should be within one standard deviation of the actual result. Therefore, a line is more accurate if s^e is a smaller number.

When Excel produces a summary output for a regression calculation, it includes various measures of the explanatory power of the regression (the following tables):

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.9781
ds, le in our estimated regression function. Because this value is very close to 1, the regression function fits our data well.	0.9567
Adjusted R Squared	0.9422
e regression line or curve. The summary regression output above indicates that the "Standard Error" is 1402.3.	1402.3
Observations The observations is the count of all of the different data points in all of the datasets.	5

For Regression, the ANOVA analysis determines whether at least one of the independent variables is significantly related to the dependent variable.

ANOVA test of significance

	[illegible] 5 - 1 = 4.	[illegible] fit.	[illegible] residual.	[illegible] 0.004.	[illegible] dependent variable.
[illegible] datasets.	1	130203045	130203045	66.217	0.004
[illegible] datasets.	3	5898925	1966308		
Total	4	136101970			

	[illegible] best-fit line of the dataset.	[illegible] dataset.	t Stat	P-value	[illegible] ("Upper 95.0%").	Upper 95.0%
Intercept	423.5	2283	0.1855	0.8647	-6841	7688
X Variable 1	10.21	1.254	8.137	0.0039	6.215	14.20

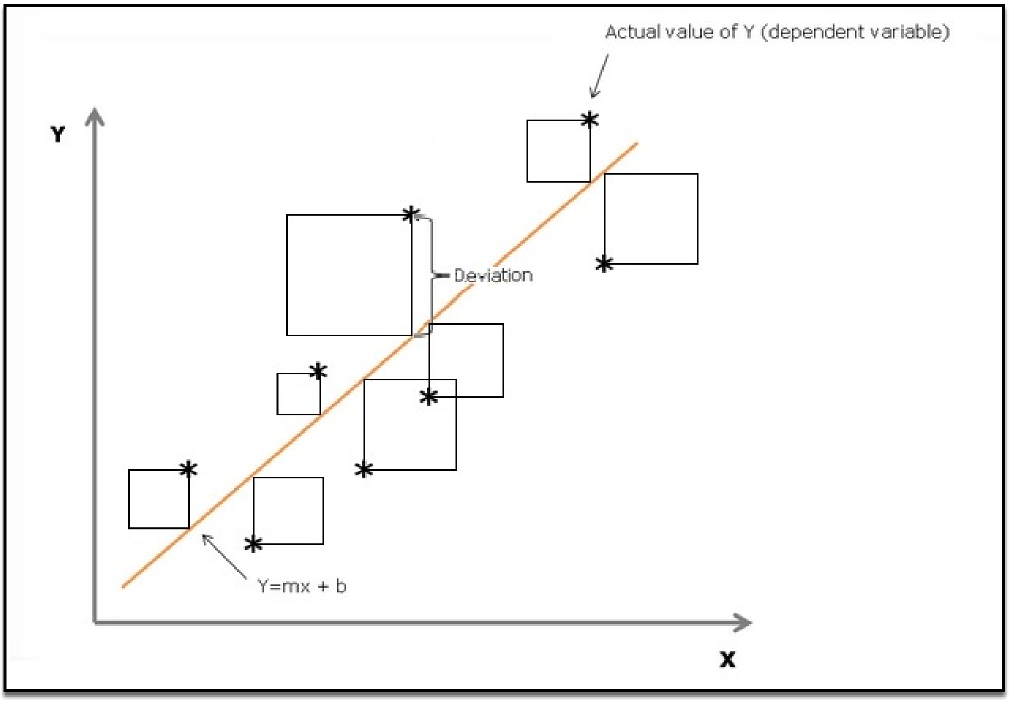
From this linear regression, the projected line will have a Y-intercept at 423.5 and will have a slope of 10.21. This creates the equation *y* = 10.21*x* + 423.5 for the best-fit line of the dataset. Therefore, if one were predicting the dependent variable value (sales revenue) when the value of the the independent variable (ad dollars spent) was $1,200, one would use the equation and determine the dependent variable would be equal to 12687.5 (*y* = 10.21 * 1200 + 423.5). Put another way, when $1,200 is spent on advertising, we can expect sales to equal about $12,687.50.

Least Squares Estimation

Most spreadsheet programs, such as Excel, can perform a least squares estimation calculation, which determines the best fit line for any of one, or multiple, variables that could affect the dependent variable of our study.

Earlier we discussed how a regression line is the "line of best fit" for a scatterplot. This is determined using an approach called "least squares." In fact, linear regression is often referred to as ordinary least squares (OLS) regression. The line of best fit minimizes the vertical distances between that line and the data points in the scatterplot. More accurately, the sum of these distances squared is minimized (i.e., least squares). The distances are squared so that points below the regression line (that will have negative distances) can have a similar contribution as points above the regression line (that have positive distances). Also, the distances are squared so that they can greater consider the points further away. This is done to minimize the distance of the line to the furthest points from it.

The following graph illustrates the concept of least squares for a number of hypothetical data points



The purpose of least squares is to minimize the sum of the squares of the differences of the results. In this graph, it shows the literal meaning of least squares. This graph is a geometrical representation of the calculation that is occurring. The "best-fit line" created by least squares is where the sum of the areas of squared deviations for all of the data points is the lowest possible sum. If one were to move the line in any way, the sum of the shown squares' areas would increase. It is the best line to represent the data.

A measure that is related to Least Squares is Standard Error of Estimate mentioned previously.

3.14 Linear Regression: An Example

Calculate the result of a linear regression equation

The regression equation, in this case, a line or simple linear regression, has the independent x variable affecting the dependent y variable. In other words, the x variables can take on any value, but the y variable is determined by an equation around the x variables. So for example:

$y = mx + b$

The y variable above is determined by the values of m and b as well as the values of the x variable. For example, the price of a bottle of wine (the dependent variable y) is a function of the demand for wine (the independent variable x): If more people want wine, it bids up the price and vice versa.

In this equation:

- b is the y-intercept, simply the value of y when $x = 0$.
- m is defined as:

  the slope of the regression line and something known as a regression coefficient. (Use the regression coefficient to say the following: "For every additional unit of x, we can predict an increase (or decrease, if negative) of m units of y)."

For example, say the best fit line, in dollars, for sales based on advertising spending is $y = 2x + 20{,}000$. If we want to calculate the predicted sales figure for spending $300,000 on advertising, we would plug in 300,000 for x, giving us the equation $y = (2 * 300{,}000) + 20{,}000$. Algebraically, we find that $y = 620{,}000$. Therefore, we would expect to have $620,000 in sales if we spend $300,000 on advertising.

3.15 Time Series Analysis

Time series analysis

Time series analysis is a technique where time is used as an independent variable to assess any influence it may have on an output. Recall that regression analysis allows for one or more independent variables, but requires a single dependent variable. Whether there is one independent variable (simple regression) or more than one independent variable (multiple regression), a time series analysis may be applicable when an independent variable represents time.

Typically, time is measured in sequential intervals of similar duration such as consecutive weeks, months or years. Multivariable regression allows for more than a single input so a test could be conducted, for example, to establish the reliability of (a) defined time intervals, (b) interest rates and (c) the unemployment rate as indicators of a dependent variable such as sales revenue.

It is risky to expect actual data following a pattern to continue to follow that pattern for inputs outside the observed data. However, it is of great interest to forecast the future so we may make decisions today that will be beneficial under future conditions that we could predict. Further, such forecasts may be quite effective predictors.

Consider the case of a business that sells school notebooks. Here seasonality is likely to be one good indicator of monthly sales since a substantial portion of units-sold annually should be delivered by the beginning of July for display during the Back-to-School (BTS) period. The percentage for BTS sales of annual sales may well be reliable. However other economic variables could influence the annual sales and thereby impact units sold for BTS even though the percentage of BTS to annual sales forms a predictable pattern.

As usual, we need to discuss a measure of error in these predictions. Complicating this is the concept of confounding variables where it may appear the independent variable studied has a direct impact on the dependent variable being predicted. Sometimes there is a third variable known as the confounding variable which directly effects the independent and dependent variables thus creating an illusion of the relationship between the two variables being analyzed that does not exist.

In a time series model, we analyze the past behavior of the time series variable in order to predict its future behavior. A graph of this past data will show whether or not there is any systematic behavior in the time series variable that would be useful in forecasting. Once data points are graphed with time as the x value, a curve can be imagined running through the points that may reveal one or more patterns that are summarized in the chart below.

Data Patterns	
Trend	A general slope upward or downward over a long period of time.
Cyclicality	Repetition of up (peaks) or down movements (troughs) that follow or counteract a business cycle that can last several years.
Seasonality	Regular pattern of volatility, usually within a single year.
Irregularity	One-time deviations from expectations caused by unforeseen circumstances such as war, natural disasters, poor weather, labor strikes, single-occurrence company-specific surprises or macroeconomic shocks.
Random Variation	The variability of a process which might be caused by irregular fluctuations due to chance that cannot be anticipated, detected, or eliminated.

Viewing past, or actual, data to construct projections allowing us to make reasonable decisions regarding the future is an ongoing interest to most of us. For example, with a trend upward, a company might ramp up production and hire more personnel in anticipation of growth, while with a trend downward, it might prepare for contraction and revise its strategy or, even, in extreme cases, exit a business.

Cyclicality and seasonality indicate to companies how to temper their expectations at certain times and to go for the gold during others. Building up inventory during slower times accommodates demand during the busiest times.

The following data can be subjected to time-series analysis. It is eight quarters' worth of revenue figures for Wal-Mart Stores Inc. from Yahoo! Finance:

Example

Wal-Mart Revenues Fiscal Quarter Revenue (in Billions)

Fiscal Quarter	Revenue (in Billions)
Q2 2004	$62.6
Q3 2004	$62.5
Q4 2004	$75.2
Q1 2005	$64.8
Q2 2005	$69.7
Q3 2005	$68.5
Q4 2005	$83.0
Q1 2006	$70.9

The data can be assessed against the following pattern types:

Trend: The data set seems to trend up.

Cyclicality: There is not enough data to determine if there is a cyclical pattern.

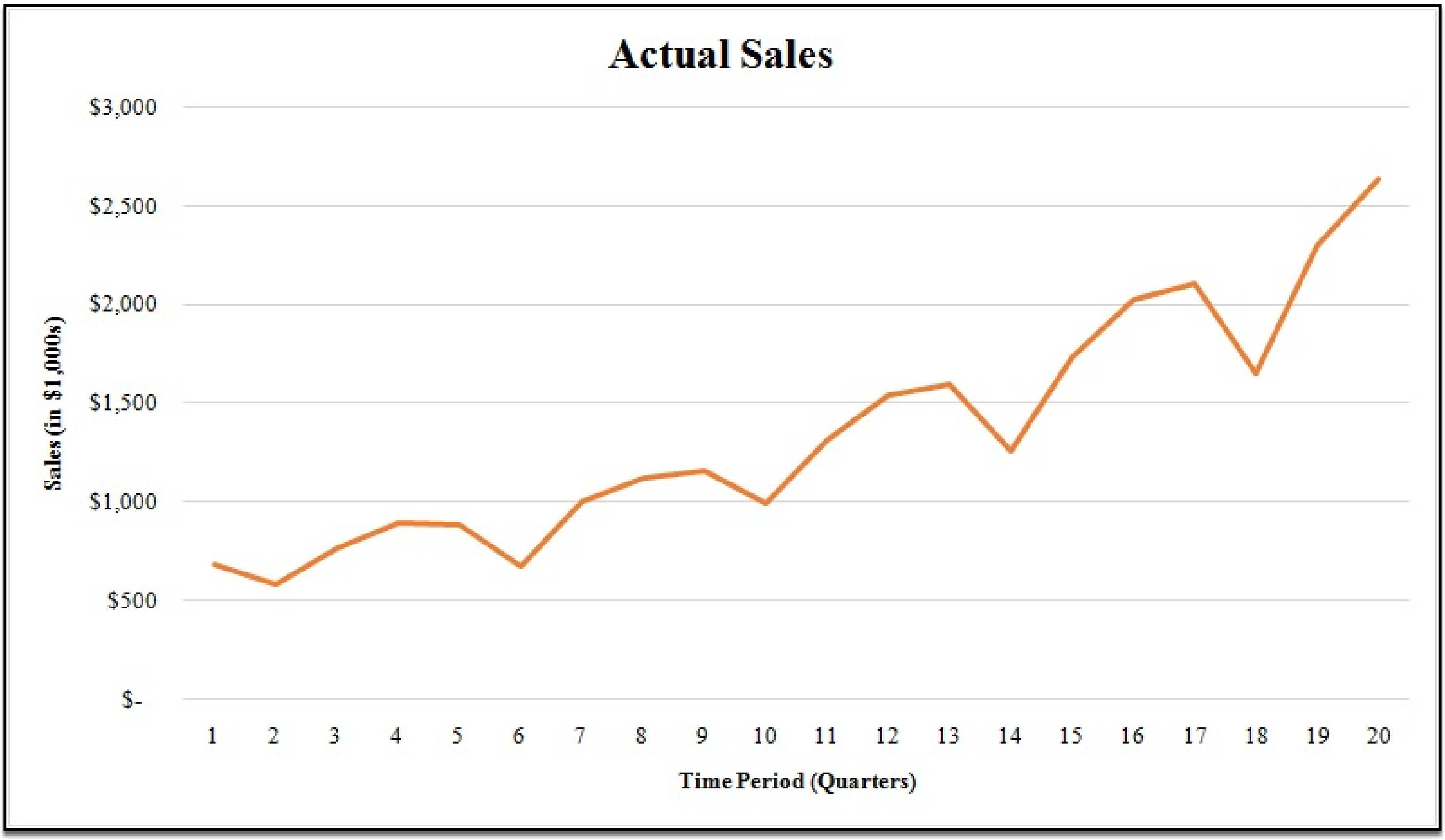
Seasonality: This data can be viewed as seasonal with a slow 1st quarter, rising 2nd quarter, slight decline 3rd quarter and a rapid rise 4th quarter. Retail, for example, books the most sales around the holidays.

Irregularity: Nothing in Wal-Mart's revenue figures seems abnormal over this brief period.

Trend Model Example

As indicated above, "trend" refers to the general direction of movement in a time series. It reflects the net influence of long-term factors that affect the time series in a fairly consistent and gradual way over time. Simply put, it reflects changes in the data that occur with the passage of time.

Consider a company that manufactures All Terrain Vehicles (ATVs). Throughout its first five years of operation, the company has enjoyed fairly steady growth in sales of its products. The officers of the company are preparing sales and manufacturing plans for the coming year. A critical input to these plans involves a forecast of the level of sales that the company expects to achieve. Quarterly sales data for the company during the past five years is provided below:



The plot of the data suggests a fairly strong upward trend in the data over time. There are a number of techniques for forecasting the value of this time series variable. One method is the linear trend model, where we would use time as the independent variable in a regression model.

3.16 Challenges with Regression Analysis

Challenges with Regression Analysis

Multiple Independent Variables

Previous assignments introduced basic concepts of simple linear regression by considering a case involving a single independent variable. Although this is appropriate in some situations, decisions in the real world often involve more than one independent variable. When multiple independent variables are involved, multiple linear regression can be used.

We have seen that creating a scatterplot between two variables visually displays what type of relationship, if any, exists. This technique does not work as well when a multiple linear regression is more appropriate. If a dependent variable is affected by more than one independent variable, creating a scatterplot based on only one of those variables might not display a strong relationship. The relationship becomes stronger, and a line will fit the data better when we account for more of the variables.

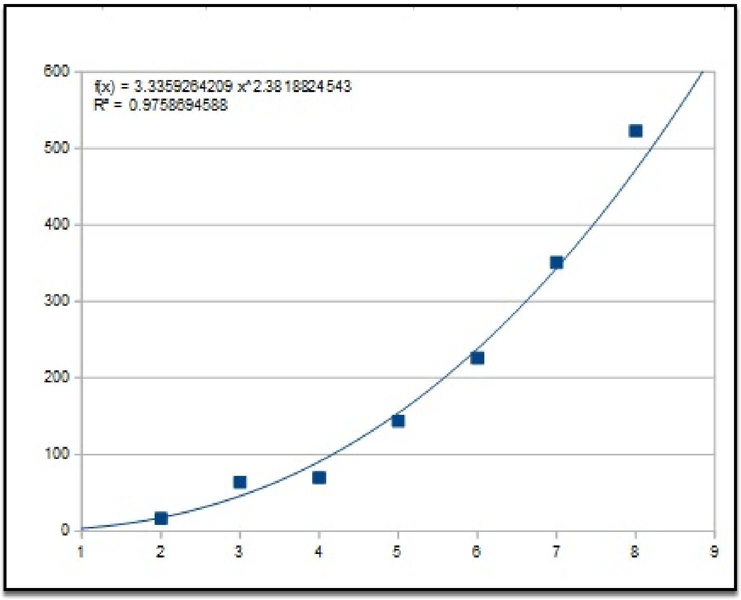
A simple linear regression model can be created for each of these variables to determine the relative strength of each relationship, but it does not paint the entire picture as well as a multiple linear regression model. For example, we can model stock prices for companies based on a variety of factors, including earnings per share, dividends per share, or stock prices of competitors. Creating a simple linear regression for any one of these variables might not yield a very strong relationship, but creating a multiple linear regression based on many of these variables may yield more meaningful results.

Non Linear Relationships

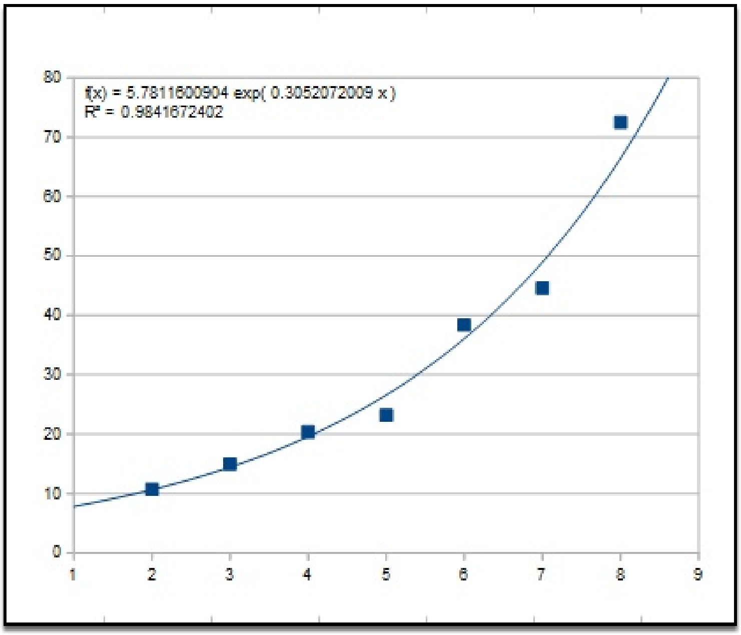
Using regression analysis, it may be the case that there is no relationship between or among variables, or perhaps there may be a relationship. Further, if a relationship exists, it is not necessarily a linear relationship. Thus, our interest in non-linear relationships.

Select any function, and we may find that function to be a good fit for some data set. Thus, data could approximate a higher-order polynomial, exponential, logarithmic or other equation. The following are some scatter plots that have datasets which would have a better fit with the adjoining non-linear functions:

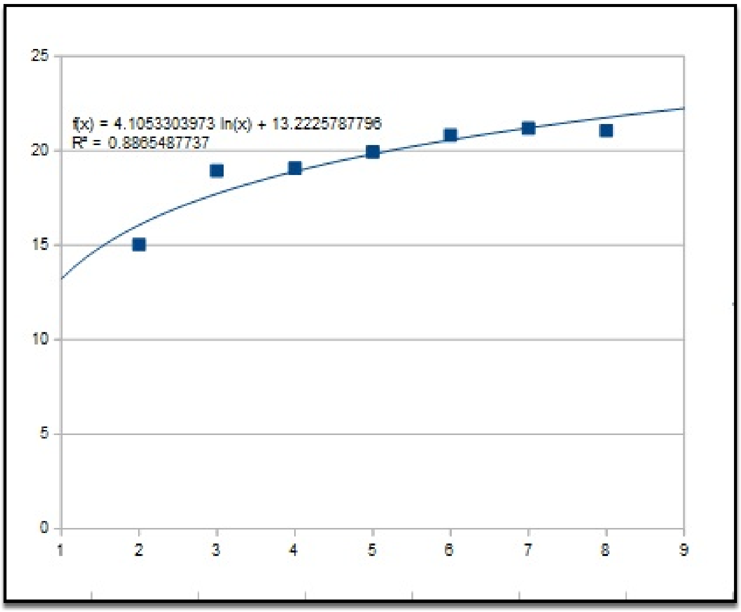
Example of a Polynomial



Example of a Exponential



Example of a Logarithmic



Outliers

An outlier is an observation point that is significantly distant from the other observations. This could be a valid data point representing a highly atypical measure, or it could be a measurement error. Either way, it is likely to cause the data set to be skewed. In regression analysis, it is important to look for outliers as they can have a big impact on the analysis. If you want to ignore these outliers, one possibility is to determine if the outlier is greater than Cook's Distance and therefore determine if it's possible to disregard it. Because outliers can represent valid and severe potential outcomes, it is important to determine the source of an outlier.

Enhanced Content

Multiple Linear Regression

Multiple linear regression is an analysis of how multiple independent variables affect one dependent variable. Regardless of the number of independent variables, the goal in multiple regression analysis is the same as in simple linear regression: to find a best fit line within our data and use it to make predictions.

If a dependent variable is affected by more than one independent variable, creating a scatterplot based on only one of those variables will likely display a weaker relationship than a scatterplot consider more than one of these variables. The relationship becomes stronger, and a line will fit the data better when more of the variables are accounted for. Separate scatterplots can be created for each of these variables, to determine the relative strength of each relationship.

Calculating Multiple Regression

A multiple regression equation supports more than one independent variable.

Example

If sales were affected by several factors, such as ad dollars spent (x_1), number of salespeople (x_2), coupons accepted (x_3), inches of rain that fell that month (x_4), etc., then each xi would have its own coefficient. So the equation would then be $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \varepsilon$. If the equation were $y = -332.5 + 10.19x_1 + 4750x_2 - 100x_3 - 825x_4$, then we could say that:

If all of the factors (explanatory variables) mentioned in this equation had their values set to zero, sales would be negative (-$332.50). For every additional $1 spent on advertising, $10.19 in sales was generated. Each additional salesperson hired earned an average of $4,750 in sales. Each coupon accepted reduced sales by $100. For every inch of rainfall during the month, the company lost $825 in sales.

How would we use the regression with four explanatory variables to predict the sales figures for a certain month? Suppose there was $1,225 spent on advertising, three salespeople selling the product, six coupons redeemed and 7.75 inches of rainfall during the month. Plug in the given values for each x as follows:

$y = -332.5 + (10.19 \times 1225) + (4750 \times 3) - (100 \times 6) - (7.75 \times 825) = \$19,406.50$

Logistic Regression

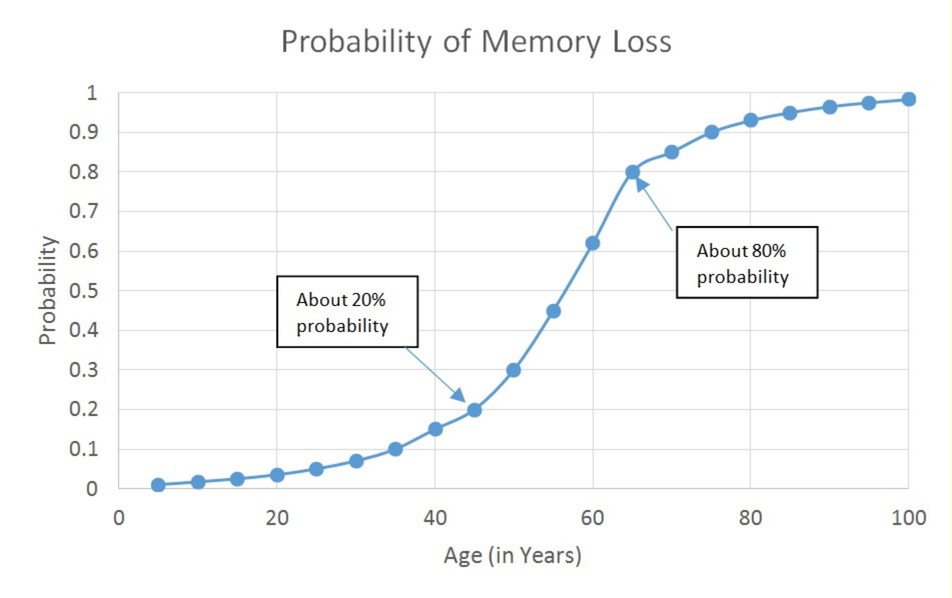
Logistic regression can be applied when the dependent variable is a categorical, binary variable, such as male/female, dead/living, gas/electric, etc. A set of data can follow a trend without that trend being linear. The data points can follow any nonlinear trend, such as a curve or series of curves. Like linear regression, logistic regression can be based on one or more independent variables. Using similar techniques like linear regression, we measure how closely our data follows a logistic regression.

Example: Tracing Memory Loss

Consider a hypothetical study of memory loss in people 65 and older. We want to estimate the prevalence of the condition as a function of age. For each person, the outcome is whether a physician diagnoses clinically significant memory loss. The predictor is the age of the patient at the time of study.

If we look at age as a predictor of the probability of memory loss, the linear regression model approach fails to provide probabilities between 0 and 1. Instead, we use the logistic function to determine probability.

Let's look at the logistic curve for this relationship. If age was strongly related to the probability of memory loss, the conditional probability of memory loss would increase as age increased, but at some point the curve flattens, indicating that increased age doesn't substantially increase the probability of memory loss. As you move downward toward a low age, the probability of memory loss might quickly approach 0, but then the curve once again flattens out. The probability of having memory loss might be similar if someone is 15 or 30, but it might make a large difference in probability if someone is 62 or 65.



As the graph shows, there is a relatively narrow age range where the probability of having significant memory loss makes a large difference, and outside of that window, the probabilities don't change a great deal. In this example, when a person reaches age 65, the probability of having memory loss is about .80. Conversely, at age 45, the probability is about .20. Beyond these points, the curve flattens out indicating less change in probability despite larger changes in age.

Autocorrelation

Autocorrelation occurs when a given data point on a time series analysis is affected by a previous data point for that time series.

In ordinary regression analysis we assume that errors are independent of one another. To determine if autocorrelation is present in a regression analysis, we can determine the Durbin-Watson statistic of the dataset. It is important to remember to check for autocorrelation when the dataset under analysis is a time series.

An example could be measuring the likelihood of snow on a given day. If the previous day was sunny and hot, it is not very likely it will snow that day. It is more likely if there was snow on the previous day. Geophysical time series are often autocorrelated.

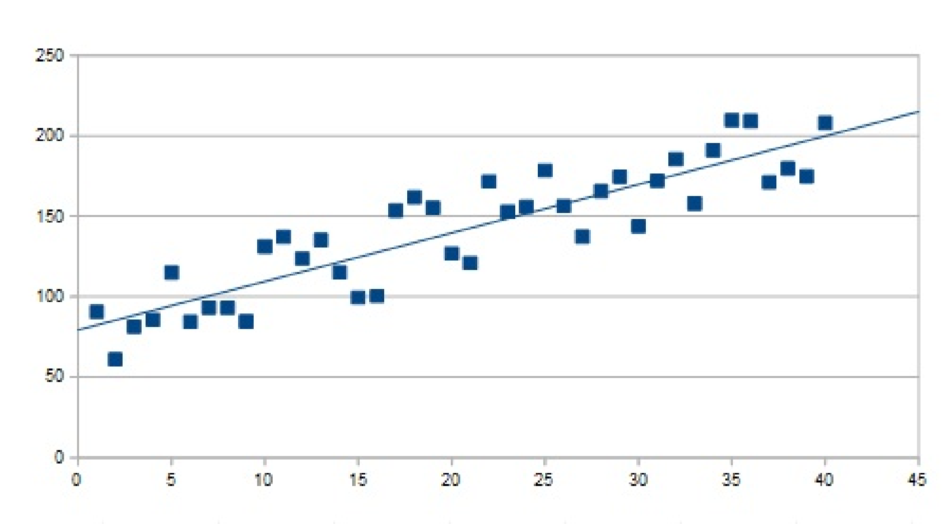
A different example could be in the business sector where a recent increase in confidence of the economy would more likely be followed by a noticeable increase in investing than an economy that had little public confidence.

Autoregressive Error Correction produces a superior regression analysis compared to ordinary regression analysis because it takes autocorrelation into account.

Homoscedasticity and Heteroscedasticity

Homoscedasticity occurs when all of the random variables have the same general finite variance. In other words, the data points in a scatterplot stay approximately the same distance away from the regression line throughout the entire dataset. For ordinary least squares (OLS) regression analysis, homoscedasticity is assumed.

Homoscedasticity Example



When heteroscedasticity occurs, the random variables have an unequal spread of variances. In other words, the data points in a scatterplot tend to be spread to varying distances from the regression line depending on the location of the data point on the line.

3.17 Cluster Analysis

Cluster Analysis

Cluster analysis, also known as segmentation, is the process of arranging terms or values based on different variables into "natural" groups. Most often with cluster analysis, these terms or values are survey responses from people. There are hundreds of approaches to cluster analysis, and it is used in many different fields to have a better understanding of an industry's environment.

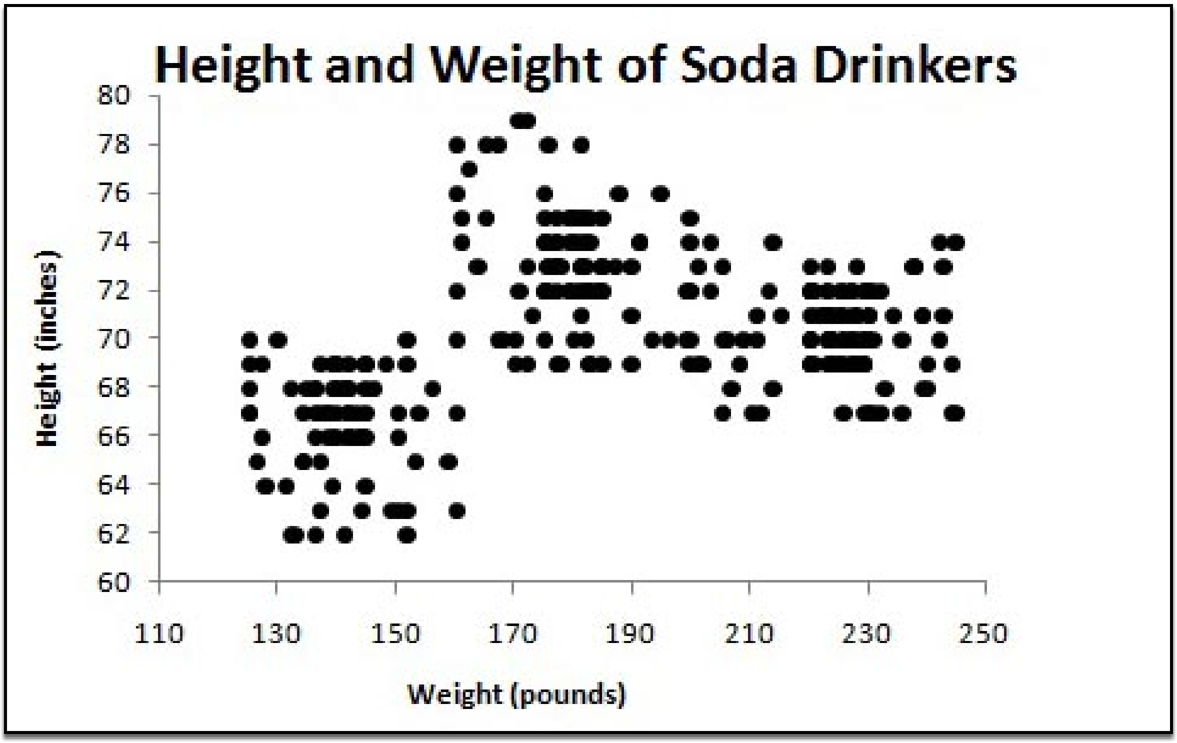
Marketing	determine target markets and for placement of products
Medicine	determine the locations of diseases by clustering differences in body makeup
Social Media	group together communities of people to create more relevant search results for certain groups
Education	identify habits and systems in place at successful or failing schools so that they can implement the working methods and get rid of the faulty methods

Example

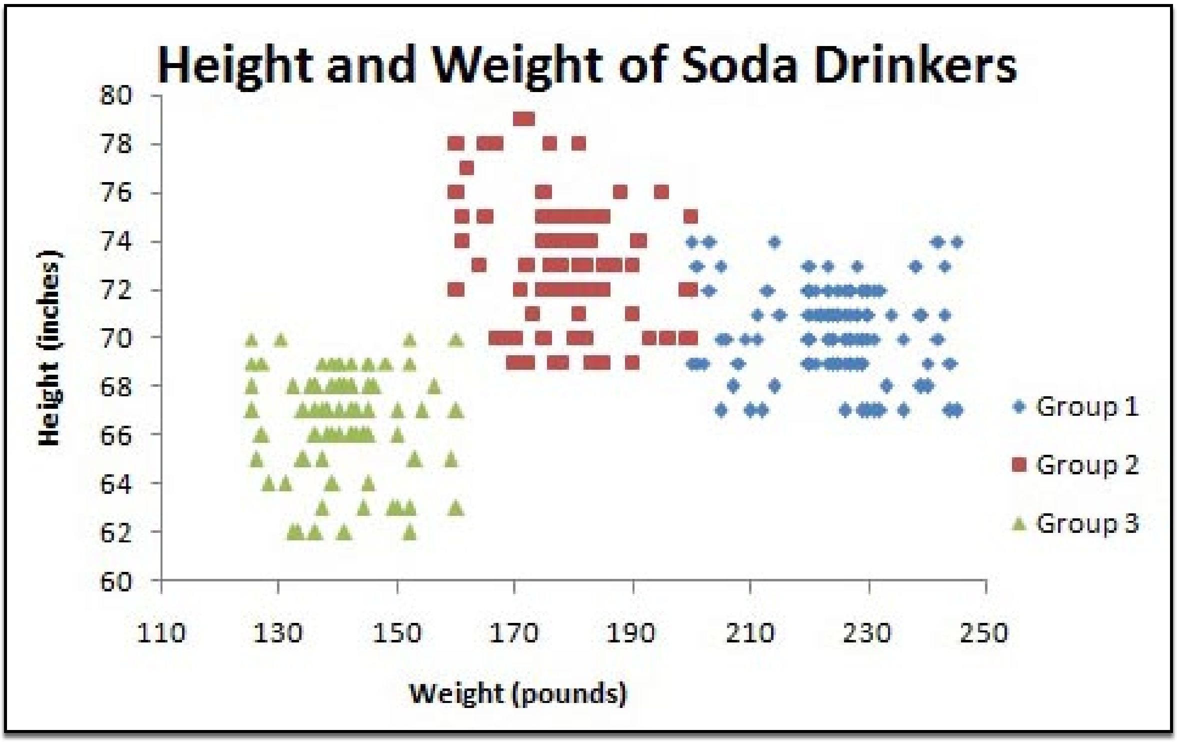
Let's imagine Emily is a marketing executive for a new soda line. Emily is trying to determine the demographics of her target markets as she conceptualizes the company's new TV commercial.

She takes a sample of 1,000 people and determines that 300 of them drink soda. From that 300, she determines the different groups of consumers she wants to target. She considers the heights and weights of the consumers.

These are the results of her data:



After applying cluster analysis, it has sorted the previous data into the following three groups:



There were 120 people in Group 1, 100 people in Group 2, and 80 people in Group 3. From this cluster analysis, Emily has determined that there are three groups. She decides that there will be three actors in the commercial, one representing each cluster.

Optional Enhanced Content

Further Cluster Analysis

Cluster analysis has a number of steps:

1. **Determine the problem by selecting the variables you want to study.**The first step is to define the problem by the variables that are being affected. For example, if the "problem" is to determine the audience of a certain television show, a network might choose to study different age and gender (these are the two variables) groups and study the viewership from each of these groups. This will help define that audience. These variables are often different demographics.

2. **Select a distance measurement between the values (there are a number of options here).**The possible methods for this are the Squared Euclidian distance, the Manhattan distance, the Chebyshev distance, and the Mahalanobis distance.

3. **Decide on which clustering procedure to use (this is where a large variation in approaches to cluster analysis occurs).** There are many different cluster procedures, but some are used more often than others. There are two main categories of cluster procedures: K-means clustering and Hierarchical clustering. K-means clustering is used when you are sorting the clusters into groups of values where the values are closest to their clusters average location. The other main category of cluster procedures is Hierarchical clustering. This is used when there is a building of a hierarchy of clustering. An example would be biological classification. A low-level in the hierarchy could be when house cats and tigers are in two different genera (low-level cluster) but are in the same family (higher-level cluster). Relative distance in variables does determine when the branches of these clusters split. Distribution-based and density-based clustering are two other categories of clustering that can be used. There are programs and Add-Ins for Microsoft Excel that can determine these types of clusters. It is important to remember that different cluster procedures are applied for different datasets and that there is no one cluster procedure that works for every dataset.

4. **Decide on the number of clusters.** If there is no pre-determined number (k) of clusters, a rule of thumb is to take the square root of half the number of data points ($k = \sqrt{\frac{1}{2}n}$). A data point is a single piece of data, usually representing an individual that has a certain answer to a question (people that drink at least x amount of soda, insects that fly, people that watch a certain tv show, trees that grow to at least x feet, etc.). The variables (possibly in the x-, y-, and even z-axes) would be dependent on the individual's responses to other questions (age, ethnicity, weight, level of education, origin location, type of tree, etc.). There are also a number of other ways to determine the number of clusters.

5. **Map the values into each cluster.**The fifth step is to display the values in each cluster.

6. **Make conclusions about the clusters.** This is where the analysis of the data is put into real-world context.

7. **Determine the validity and reliability of the analysis.**The validity can be tested by using different distance measurements (step 2), randomly splitting the data in half and analyzing the results for each half, or repeating the analysis several times and either deleting a variable each time or using a different order each time.

3.18 Decision Analysis

Decision Analysis

Decision Analysis is the process of weighing all outcomes of a decision to determine the best course of action. This is for any situation a person faces, whether personal or professional.

There should be an analysis of the current situation whenever making a decision. The first piece of the process is to define the situation and determine the different possible actions. From this, it's important to determine the possible outcomes of those actions. The next step is to assign costs and benefits to each action and outcome. Now that you have the value of each action and outcome, it is important to determine the worth of each action and choose the best-expected outcome.

There are three different contexts in which one can make decisions.

Certainty

Under certainty, each action has known outcomes. The outcomes are weighed, and the best action should be chosen.

An example could be if Stephanie had a number of house offers on her table. She is selling her house, and when she accepts any one of them, it's sold. All that has to be done is to determine the best offer.

Risk

Under risk, there are known or estimated probabilities of the outcomes of different actions. Knowing the probabilities of the different outcomes allows for the calculation of an actions' worth. To calculate the expected payoff for an action, you multiply the outcomes' payoff by the probability of that outcome; then all of these together for each outcome. The action with the greatest expected payoff is the option chosen.

For example, in the decision of the town fair or the night at the movies, let's assume there's a 70 percent chance of rain.

The expected payoff for the fair is

P(F) = 0.7 × (- $2,000) + 0.3 × (16,000) = $3,400

While the expected payoff for the movies doesn't change because the rain doesn't affect the movies.

P(M) = 0.7 × $5,000 + 0.3 × $5,000 = $5,000

A decision tree can be especially useful in determining outcomes under risk.

The Decision Tree

A more-complex quantitative analysis approach involves a tool called a decision tree.

A decision tree is a decision analysis tool that shows a number of options, the paths by which each of these options may be reached, and the possible consequences of choosing each option. A decision tree analysis is designed to establish a logical sequence for decisions, to consider the decision alternatives available, and to evaluate the results they will produce.

Decision trees are valuable for risk assessment when a limited number of potential outcomes are possible. Decision trees use the weight of each risk and estimates to determine the potential impact for various alternatives. When there is a point in a project where several options are possible, practitioners analyze each option and assign it a probability. The probabilities for all the options should total 100%.

Typically decision trees are used with risks that affect the project schedule or project costs. Through a decision tree analysis, project participants can select the option that has the lowest possible impact.

Decision trees can be applied using, or not using, probabilities. One technique with probabilities is the expected value approach. (Non-probability techniques include minmax regret, maximum regret, and maximum payoff.)

Decision trees can help select the best course of action in situations of uncertainty.

Decision Tree Example

Here is an example of how a decision tree analysis is performed.

Suppose a luxury car dealer must decide how many European sports cars to order for the coming year. It costs her $25,000 to keep each unsold car in inventory, and her gain (profit) is $50,000 on each car sold.

Further, suppose that the car dealer can order 0, 1, 2, or 3 cars. Here is her decision table, showing the potential pay-offs for any given course of action.

Actual Annual Demand	Decision: Order 0 Cars	Decision: Order 1 Car	Decision: Order 2 Cars	Decision: Order 3 Cars
0 Cars	0	($25,000)	($50,000)	0 - 3 × $25,000 = ($75,000)
1 Car	0	$50,000	lered. 1 car sold, 1 car 2 1 × $50,000 - 1 × $25,000 = $25,000	0

2 Cars	0		$100,000	$75,000
3 Cars	ars cars	$50,000	$100,000	$150,000

1 × $50,000 - 0 = $50,000

unsold. 0 - 0 = 0

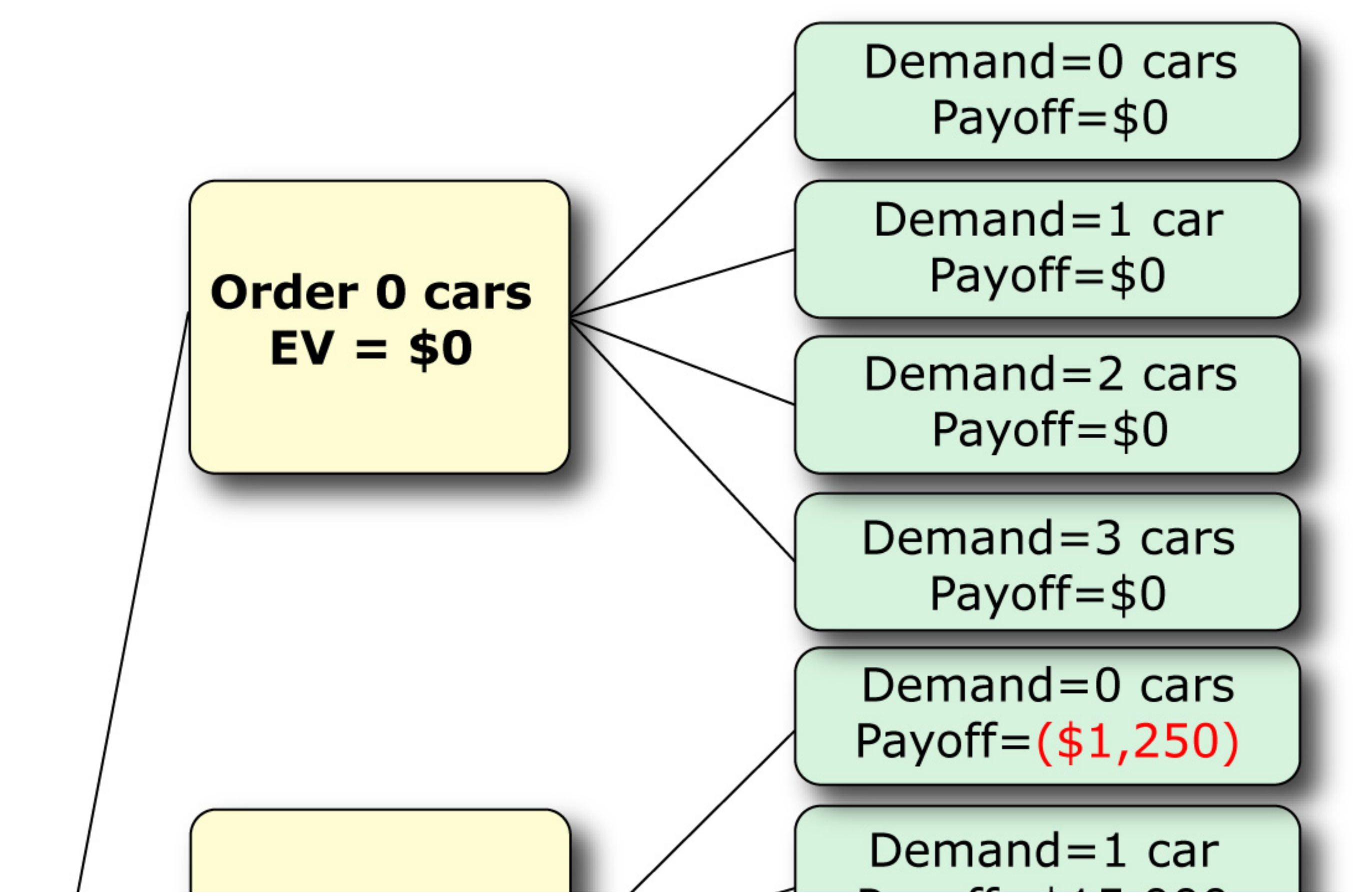
How can this car dealer figure out how many sports cars to order? She looks at market research and historical trends to develop a table of the probabilities of the annual demand for the cars.

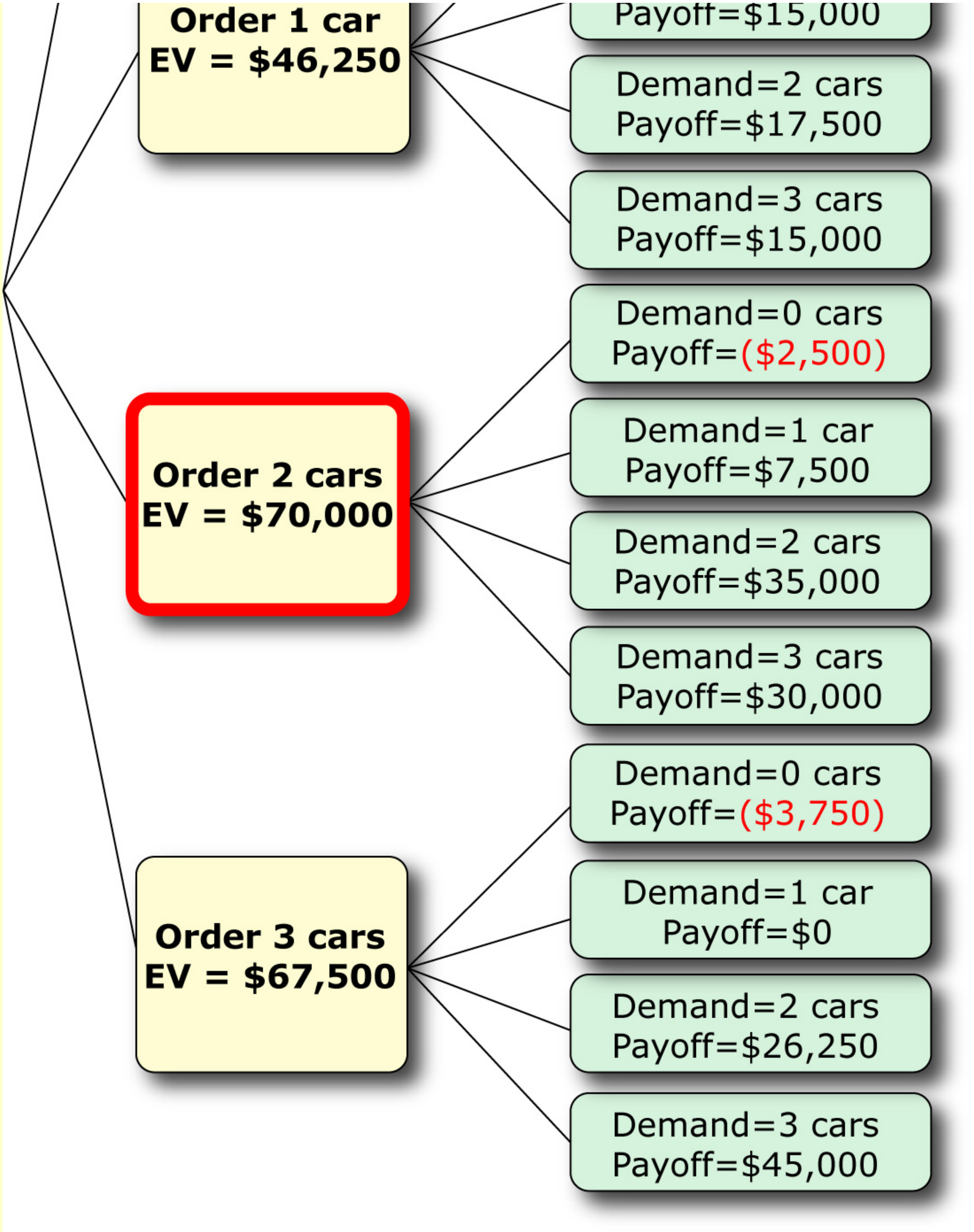
Probability	Demand
.05(5%)	0 Cars
.30(30%)	1 Car
.35(35%)	2 Cars
.30(30%)	3 Cars

Combining the probabilities with the potential pay-offs allows the car dealer to come up with the**expected value** of each decision.

Projected Annual Demand	Decision: Order 0 Cars	Decision: Order 1 Car	Decision: Order 2 Cars	Decision: Order 3 Cars
0 Cars (5%)	0	5% × $25,000 = $1,250	5% × $50,000 = $2,500	5% × $75,000 = $3,750
1 Car (30%)	0	30% × $50,000 = $15,000	30% × $25,000 = $7,500	30% × 0 = 0
2 Cars (35%)	0	35% × $50,000 = $17,500	35% × $100,000 = $35,000	35% × $75,000 = $26,250
3 Cars (30%)	0	30% × $50,000 = $15,000	30% × $100,000 = $30,000	30% × $150,000 = $45,000
	EV=0	EV of 1 Car = $46,250	EV of 2 Cars = $70,000	EV of 3 Cars = $67,500

The following is the decision tree. The payoff is the probability of the outcome multiplied by the loss/profit of that outcome.





Based on these expected values, the car dealer would order 2 sports cars; that is the decision with the highest expected value.

Warning

With uncertainty and risk, it might not be a good decision to follow this decision analysis and go with the maximum profit or maximum expected payoff. This is important because it is possible that, although those options might theoretically give the highest average outcome, the losses that could be involved might be too great that the action might not be worth taking. This is a judgment call, but it is important during decision analysis to be aware that the maximum loss might occur. The expected payoffs would be the best choice if done thousands of times but might not be worth it in a situation if done once where the one-time loss could have a huge effect.

3.19 Simulations

Simulations

A simulation attempts to emulate a real process or system through an imitative model. This allows considering problems that may not lend themselves to direct experimentation and helps managers make decisions. Common simulation tools include what-if analysis, and Monte Carlo simulation.

Simulation models use computer-based programs to predict the behavior of a system. The simulation acts as a similar but simpler model to represent a situation or problem in order to analyze possible outputs. It will include observed data as well as randomly generated, predicted data.

Simulations can help in testing the sensitivity of different changes to a system, in assessing risk, in reviewing possible outcomes from differing actions, and to help in uncovering patterns.

Mathematical expressions are used to describe the relationship between inputs and outputs in a simulation model. There are two types of inputs:

- Controllable inputs: Inputs that are directly controlled by the company or individual. An example would be the decision by a company to invest in building a new production facility.
- Probabilistic inputs: Inputs that are outside the direct control of the company or individual and take on different values. An example would be the costs of raw materials used in the new production plant the company had decided to build.

Simulations are helpful for creating summary statistics for a real-world process. For example, we might predict the following probabilities in a given week for workers' compensation claims:

Number of claims in a week	Probability of this number of claims in a week
0	0.80
1	0.15
2	0.04
3	0.01

We can use these predictive statistics figures as a model to run a simulation. Depending on preference and relevance to the situation, we can run a simulation for whatever period of time makes the most sense; let's say 52 weeks, here. Using a random number generator and the probabilities stated above, we can see after a simulation of 52 weeks what some possible totals for worker's compensation claims could be in the following year.

Discovering an unexpected simulated outcome can help us better prepare for volatility in professional settings. A simulation can show us that a contingency plan is necessary to prepare for a certain number of malpractice suits in the medical field, a fundraising level in a nonprofit, a frequency of equipment failure in the manufacturing field, and many more. To be well prepared for any outcome, professionals will often use simulations to give them an idea of what could happen.

What-if Analysis

What-if analysis is a form of simulation analysis that involves selecting different values for the probabilistic inputs in a model and then computing the possible outputs.

For example, a company could use "what-if analysis" to model the outcomes (output) caused by altering the demand (a probabilistic input) for a new product. Or a company could assign a probability distribution to those inputs and then run the model with those new variable values and look at the output.

Example: The CEO of a construction firm is curious about what will happen to profitability if the sales of new housing units increased, but the price of lumber and labor expenses went up. A "what if" model could look at the sensitivity of profitability based on differing inputs.

Monte Carlo Simulation

A Monte Carlo simulation is a problem-solving technique used to approximate the probability of certain outcomes by running multiple trial runs, or simulations, using random variables. It lets us model situations that present uncertainty and run them thousands of times on a computer.

This probabilistic model can be used to find an approximate solution to a numerical problem that would be difficult to solve by other methods. Some firms use the Monte Carlo method to estimate the riskiness of new products, to price complex financial derivatives, or to model foreign exchange risk and find the best way to hedge that risk. Financial planners use Monte Carlo simulation for retirement planning to determine optimal investment strategies.

While there are several different ways to run a Monte Carlo simulation, there are five basic steps:

1. Decide the probability distribution of important variables.
2. Calculate the cumulative probability distribution for each variable.
3. Decide an interval of random numbers for each variable.
4. Generate random numbers.
5. Run a series of trials and determine simulated value of the actual random variables.

Example: A company needs to assign a price to stock options before the end of the month. A Monte Carlo simulation can generate thousands of possible (but random) price paths in order to estimate the future option value. (Options change in value based on the price of an underlying stock, which itself can be affected by a very large number of factors.) The future option value can be used to assign a current option price.

Simulations: Advantages and Disadvantages

Simulations have advantages and disadvantages, as summarized in the following chart.

	Advantages and Disadvantages of Simulations
Advantages	• allows analysis of large, complex, uncertain systems • permits studying systems behavior without disruption of actual operations • can compress a time frame and enable assessment of effects of a change that take place over several years • can employ a large number of probabilistic inputs • allows the consideration of "what if" scenarios • can investigate scenarios that might prove dangerous in real life (such as modeling for healthcare or manufacturing)
Disadvantages	• can be expensive and time-consuming to develop and execute • dependent on realistic, accurate input • can require user to have deep understanding to generate conditions and constraints • can be difficult to interpret simulation results • dependent on assumptions (such as probability estimates) that may turn out to be incorrect

3.18.1 Monte Carlo Simulation in the Real World: Financial Planning

Monte Carlo Simulation in the Real World: Financial Planning

In finance, as in many other fields, there are variables whose future values involve uncertainty which can be modeled by randomness. This makes Monte Carlo methods suited for many financial problems.

If we want to find the expected value of a portfolio at some future date T, we can simulate possible paths the assets' values can take from now to then, find the values from each replication at date T and average them to get an estimate. Finding the probability of a given event can be done in a similar way.

In Monte Carlo methods, assumptions have to be made about how to model the randomness in variables affecting the asset's value. The degree of realism of these assumptions is part of what makes a particular Monte Carlo model a good or bad predictor of future outcomes.

Monte Carlo Methods and Financial Calculators

A financial calculator for retirement planning is software that employs user information (such as age, income, spouse's income, assets, Social Security benefits, and spending habits) and simulates a large number of possible future outcomes for the value of the user's wealth.

Relevant variables such as inflation, interest rates, taxes, stock returns, and life expectancy are assumed constant or assumed to follow some stochastic process (one involving random variables). The randomness in some of these variables is what makes the outcome of each replication of a simulation different.

Common questions that can be answered by a financial calculator include:

- What is the expected value of my assets at retirement (given a certain level of contributions)?
- How much do I have to contribute per year in order to maintain a certain living standard throughout retirement?
- What is the probability that my pension funds will run out before the age of 85 (or before the end of my life)?

Example

Suppose that an individual wants to find the probability that a portfolio made of 40 percent stocks and 60 percent bonds will last through 25 years of retirement, say from 2015 to 2040.

One type of Monte Carlo simulation would assume returns to be normally distributed at each future date. The parameters of the normal distributions might be chosen to match historical data. Another type of Monte Carlo simulation would take a random draw from a year from some historical period (for example, 1900 to 2000) and use the average asset returns from that year as an estimate of the returns for 2015, then take another random draw to get estimated asset returns for 2016, and so on until the year 2040. Expected tax, interest and inflation rates may be forecasted, assumed constant or assumed to follow some stochastic process. Different rates of withdrawal can also be assumed; suppose that the individual expects to withdraw \$40, 000 per year after retirement.

At each iteration of the simulation, one path for tax rates, interest rates, inflation and asset returns is obtained, and the estimated value of the portfolio each year is calculated. Ten thousand (10,000) independent replications of this simulation are made. Suppose that in 1,000 of the replications the portfolio ran out of money by 2040, and in 9,000 replications it did not. Then the estimated probability of the portfolio lasting 25 years after retirement is $0.9 = \frac{9{,}000}{10{,}000}$.

The user then has to decide what probabilities she is willing to accept. If 0.9 seems too low, she can decide to save more before retirement, change her portfolio composition to one that appears more likely to last, or accept a lower withdrawal rate after retirement. These simulations can be repeated with different assumptions on the withdrawal rate and on the other variables to give a wider range of possibilities.

Concerns about Monte Carlo Methods and Financial Calculators

The recent financial crisis and the ensuing 2008-2009 recession alerted investors to the dangers of putting too much confidence in the probabilistic predictions of Monte Carlo models, at least with the assumptions in use at the time.

In general, these methods gave extremely low probabilities for such a great loss of wealth. The financial planning Monte Carlo models under criticism assumed normal distributions for stock returns at any point in time. This assumption is made in part for greater tractability of the models; but long before the financial crisis researchers had recognized that the variation in stock returns is not well described by normal distributions. The mathematician Benoit Mandelbrot suggested that stock returns are not only not normal, but may be skewed, and proposed using stable Paretian distributions. However, these distributions have other inconsistencies with the observed behavior of stock returns, such as infinite variance.

Paul Kaplan of Morningstar conducted an empirical study of historic Standard&Poor 500 data and concluded that monthly returns fell over three standard deviations below average five to ten times more often than predicted by a normal model. A random variable that follows a normal distribution has a probability of only 0.003 of falling more than 3 standard deviations away from its mean. Kaplan claimed that a log stable model provided better probabilistic predictions. [1]

Not all experts agree that normal models should be replaced. In 2009, David Hultstrom, CFP, CFA, ChFC, pointed out that in the last 83 years of market returns data, there have been no cases of *yearly* returns falling outside of 3 standard deviations from the historical average.[2]

Instead of assuming normal distributions for stock returns each year, one could use the alternative method described in the example above, where predicted future asset prices are randomly drawn from historical values. This is approximately the method used by the Vanguard retirement calculator, as described on the Vanguard website.[3]

Another criticism of the types of Monte Carlo methods traditionally used is that they did not run enough simulations to get accurate results. It was common to run only hundreds or thousands of replications, whereas the Retirement Income Industry Association has recommended running at least tens of thousands.

In using Monte Carlo results, it is important for financial planners to explain to their clients how the inputs and assumptions affect the probabilities of different events. Financial planners should also update their clients' predictions regularly, according to Ed McCarthy, CFP. A forecast made closer to a target date may give a very different result from a previous forecast, as much more information may have become available.

Monte-Carlo Alternatives

Before the development of more powerful computers made Monte Carlo simulation feasible, financial planners were more likely to make non-random predictions. For instance, they might assume that the rate of returns on assets is constant and equal to the historical average, the inflation rate is constant, tax rates are constant, etc. Then the predicted value of the portfolio would just be a projection and could be calculated by hand.

A financial planner could run "worst case," "best case," and intermediate scenarios, in which the worst case scenario assumes a low yearly return, high inflation, etc. and the best case assumes an opposite situation. While such methods provide a range of possibilities, they do not give the probabilities of the different cases, and the range of possibilities may be quite wide so that it is difficult to make decisions based on these results.

Since the crisis, some financial planners have continued to use the same Monte Carlo methods as before, but with added provisions for very bad outcomes. They might take the results from a Monte Carlo calculator, but add more probability to the worst outcomes, and act accordingly. Some planners run simulations using several different distributional assumptions for the relevant variables, a kind of "sensitivity analysis." As Monte Carlo methods incorporating fatter-tailed distributions are developing, it will be interesting to see whether they produce more accurate predictions than previous models.

Notes

1. J.P. Bruckenstein, "What are the Odds?" *Financial Advisor Magazine*, September 2009..

2. Ed McCarthy, "Defending Monte Carlo Simulation," ThinkAdvisor, 2009.http://www.thinkadvisor.com/2009/06/10/defending-monte-carlo-simulation

3. Vanguard Retirement Nest Egg Calculator
https://retirementplans.vanguard.com/VGApp/pe/pubeducation/calculators/RetirementNestEggCalc.jsf

3.18.2 Exercise: Monte Carlo Simulation

This assignment does not contain any printable content.

3.20 Vocabulary Game

This assignment does not contain any printable content.

3.21 Flashcards

Flashcards

Term	Definition

3.22 Case Study: ShinyWhite and simple linear regression

Simple Linear Regression: Evaluating the Relationship between Advertising Expenditures and Sales

Sarah Wu is the Product Manager for the ShinyWhite toothpaste brand that is part of the consumer products portfolio of Mega Enterprises, a multinational conglomerate. To a large extent, Wu has been successful targeting ShinyWhite at younger, "beauty-conscious" female consumers, and sales have risen steadily since the product was introduced 10 years ago.

At the corporate level, however, there has been tension over the company's recent quarterly earnings reports. Mega Enterprises has missed its targets, and its performance has upset shareholders. The message coming from the boardroom is that Mega will need to rein in expenses across all divisions in order to get earnings back on track.

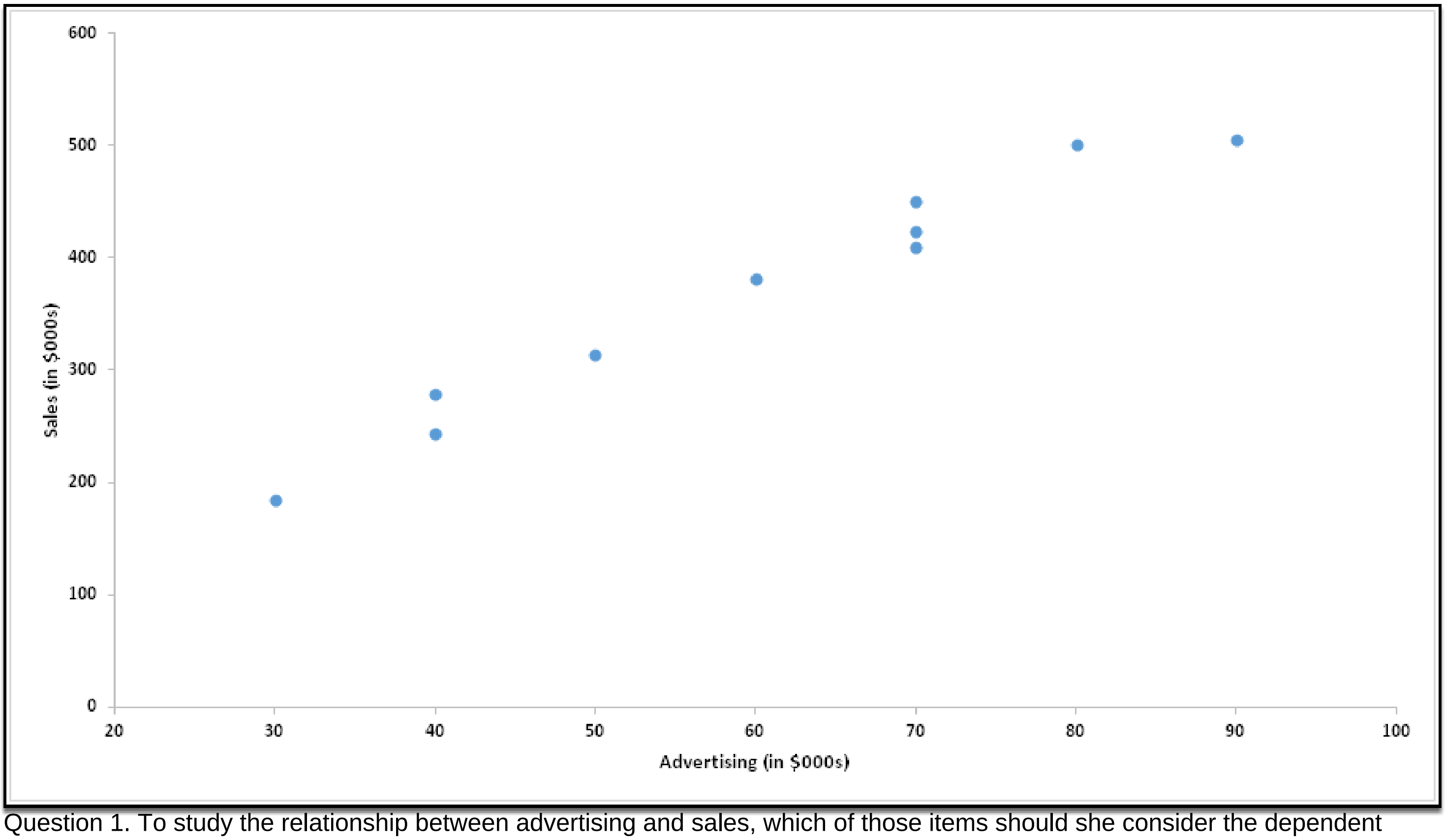
Wu is concerned that proposed cost-cutting might hurt ShinyWhite's advertising budget. She believes strongly that the successful sales growth of the product has been largely influenced by the amount of money spent on well-targeted advertising. In anticipation of proposed advertising budget cuts, Wu decides to construct a regression analysis that she hopes will demonstrate the efficacy of advertising and its direct influence on ShinyWhite's sales.

Wu also expects the regression analysis to help her predict the average level of sales that she can expect, given a particular advertising budget. She knows that her division's senior managers respond well to data-based arguments and presentations.

Wu's first task is to compile data on the level of ShinyWhite sales observed for various levels of advertising expenditures in 10 different test markets around the country. She chooses test markets that are similar in terms of size and other demographic and economic characteristics. These figures are as follows:

Observation #	Advertising (in $000s)	Actual Sales (in $000s)
1	30	184.4
2	40	279.1
3	40	244.0
4	50	314.2
5	60	382.2
6	70	450.2
7	70	423.6
8	70	410.2
9	80	500.4
10	90	505.3

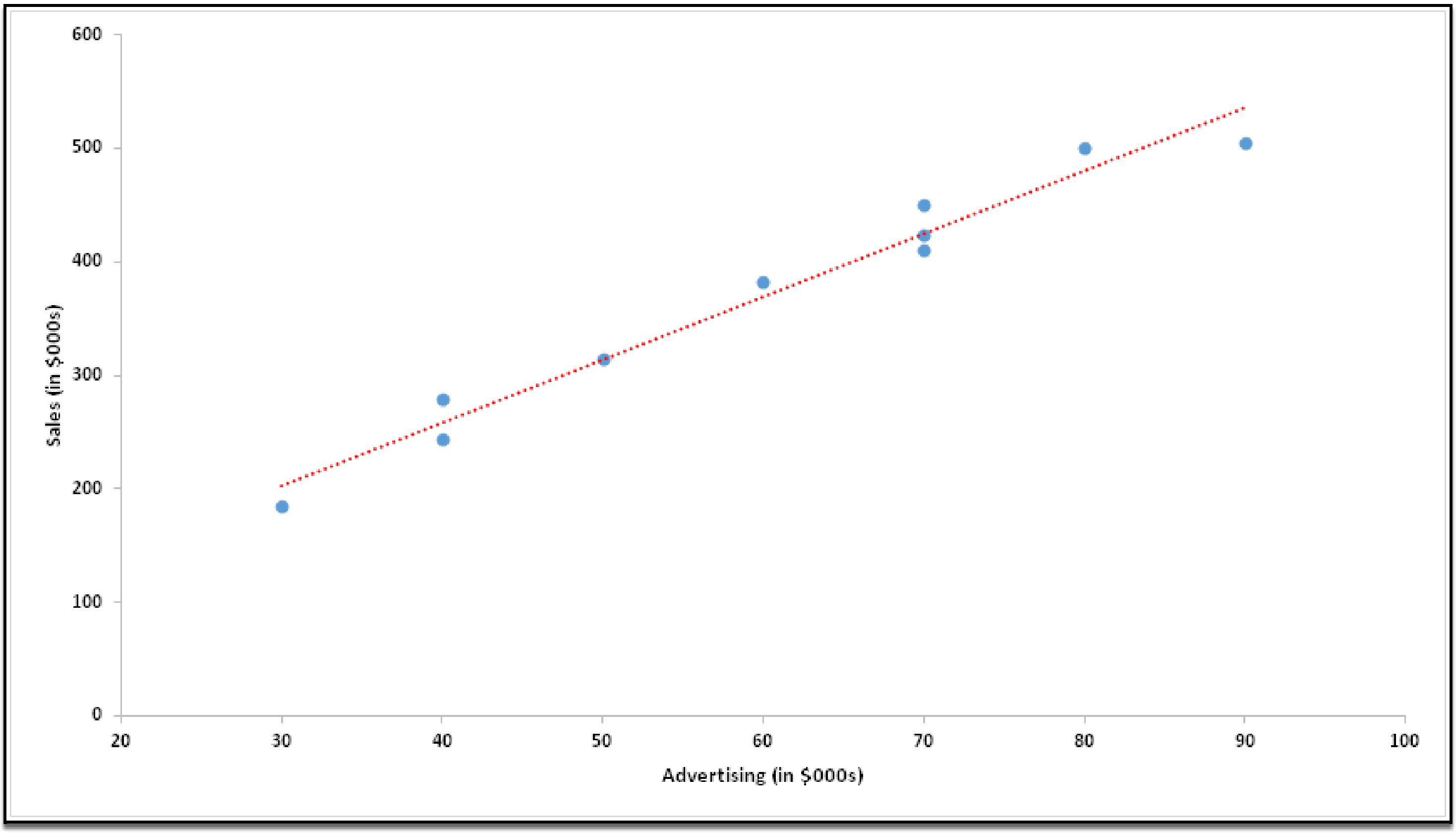
Wu then displayed this data graphically using a scatter plot to show the relationship between advertising and sales:



Question 1. To study the relationship between advertising and sales, which of those items should she consider the dependent variable (Y) and which is the independent variable (X)?

Question 2. Does the graph above suggest a strong linear relationship between advertising expenditures and sales? Is it a perfect functional relationship (where each level of advertising would yield one unique level of sales)? Why or why not?

Wu then draws a straight line through the data in the graph (see red line on graph below) and uses the standard linear regression formula y = mx + b to calculate a formula for this line that would account for the systematic variation between advertising and sales (defining "best fit" for the line that best fits the sample data).



She uses Excel to determine the values of "m" and "b" for the equation of the line that fits the data. Using the "Regression" feature of Excel's Data Analysis tool, she produces the following regression summary output, based on her data:

Regression Statistics	
Multiple R	0.984444276
R Square	0.969130533
Adjusted R Square	0.96527185
Standard Error	20.42132374
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	1	104739.6003	104739.6	251.1558	2.51425E-07
Residual	8	3336.243706	417.0305		
Total	9		108075.844		

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	36.34235294	21.98328259	1.653181	0.136894	-14.3511876	87.03589349	-14.35118761	87.03589349
X Variable 1	5.550294118	0.350222813	15.84789	2.51E-07	4.742678863	6.357909372	4.742678863	6.357909372

Residual Output

Observation	Predicted Y	Residuals
1	202.8511765	18.45117647
2	258.3541176	20.74588235
3	258.3541176	14.35411765
4	313.8570588	0.342941176
5	369.36	12.84
6	424.8629412	25.33705882
7	424.8629412	1.262941176
8	424.8629412	14.66294118
9	480.3658824	20.03411765

10	535.8688235	30.56882353

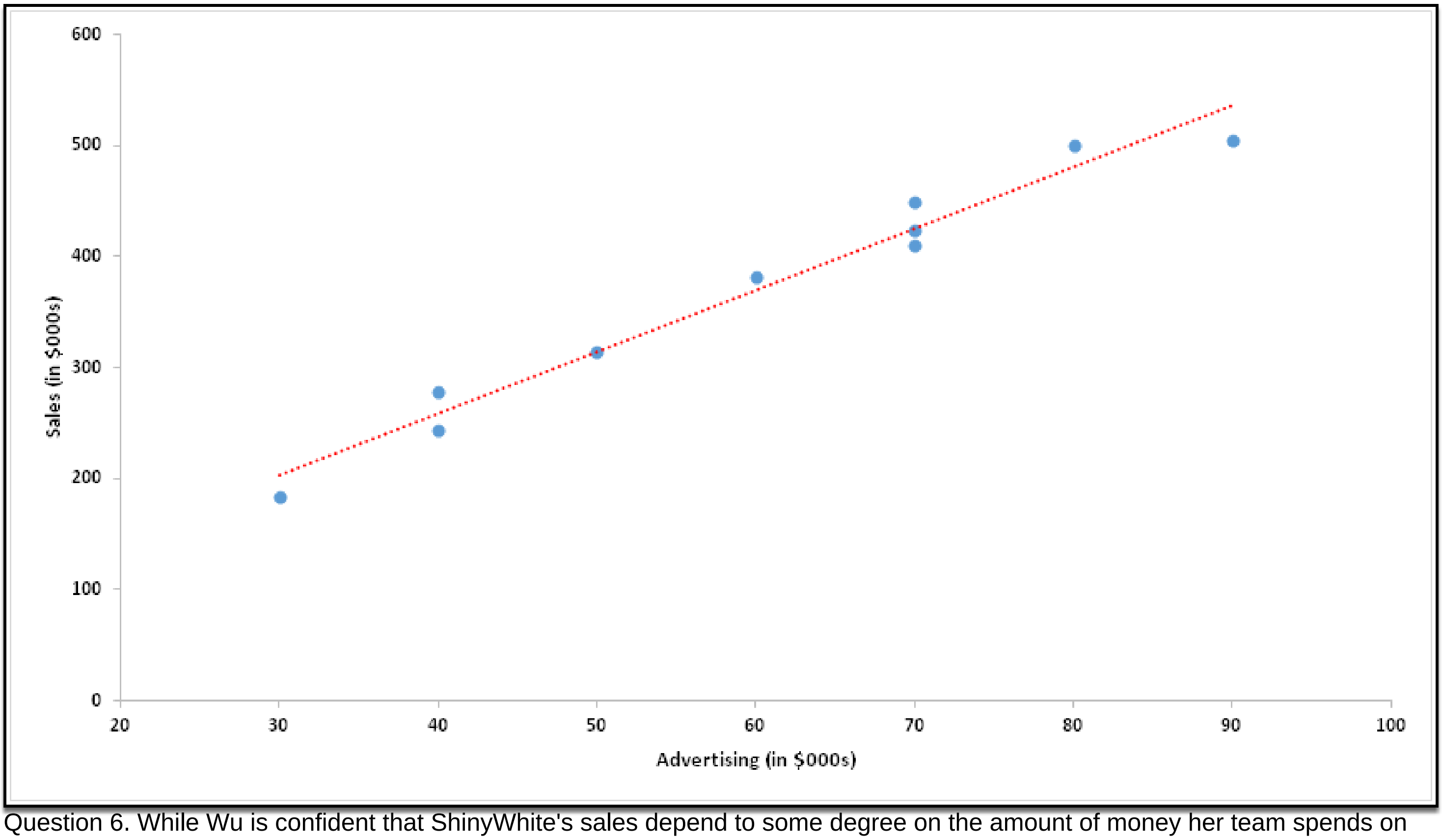
While there are many useful regression outputs in the above table, Wu located the "Coefficients" to complete the formula for the best fit line:

y = 5.55x + 36.34

Question 3. Does the graph suggest that there is a "statistical" relationship between the variables?

Question 4. Using this formula, what level of sales could Wu expect to achieve for ShinyBrite if she spends $65,000 on advertising in a given market?

Question 5. Based on the formula Wu calculated, why can't she be certain that $65,000 in advertising expenditures will produce the sales result she predicted above? What could she calculate to evaluate the level of uncertainty in her sales predictions from the regression formula?



Question 6. While Wu is confident that ShinyWhite's sales depend to some degree on the amount of money her team spends on advertising, do you believe there are other factors that might also affect the product's sales? If so, what do you think those could

be?

Question 7. If Wu wants to present her findings to Senior Management in order to protect the advertising budget levels for ShinyWhite, do you think the regression analysis and graph she prepared above is enough to make her case? If not, what approach should she take to establish better evidence of the influential link between advertising expenditures and sales?

Suggested/Sample Responses

Question 1:

The dependent variable (Y) is sales, and the independent variable (X) is advertising. The theory is that sales are dependent (to some extent, at least) on the level of advertising.

Question 2:

Yes, the graph above suggests a strong linear relationship between advertising expenditures and sales. As advertising expenditures increase, sales increase proportionately. However, the relationship between advertising and sales is not perfect. For example, advertising expenditures of $70,000 were used in three test markets and resulted in three different levels of sales. Thus the level of sales that occurs for a given level of advertising can fluctuate.

Question 3:

Yes. Each level of sales is associated with a range or distribution of possible sales values.

Question 4:

(5.55) x (65) + 36.34 = 397.09

If Wu spends $65,000 on advertising (in a market similar to those used to estimate the regression function) she could expect to observe sales of approximately $397,090.

Question 5:

The actual level of sales is likely to differ somewhat from the answer above due to other factors influencing sales. A standard error would help her determine the level of certainty, as well as a range of values for the actual sales volume (with upper and lower limits) that would occur with 95% certainty.

Question 6:

Yes, there are almost definitely other factors that affect sales. For example, general economic conditions, the level of competition in the marketplace, product quality, shelf placement at stores, etc.

Question 7:

The linear regression analysis above is helpful and shows that there is a relationship between the level of advertising and the sales. It is better than instinct alone in making Wu's case for the importance of advertising expenditures in keeping sales robust. If, however, Wu wants to make a stronger and more persuasive case, she should take into account the involvement of other independent variables that are likely to be important in explaining sales results. To better make the case that advertising is the "most" important variable, she should carry out a multiple regression analysis.

3.23 Module 3 Printable PDF

This assignment does not contain any printable content.

3.25 Module Feedback

This assignment does not contain any printable content.