

CHAPTER TWO

PROBABILITY & THE NORMAL DISTRIBUTION

Basics of Probability

What comes to mind when you hear the word *probability*? Most people think of gambling and it is understandable since you can never bet on a sure thing. But what is probability really? It is the chance or likelihood of an event occurring. Whether you are deciding to take an umbrella because you just heard there would be a 50% chance of rain or you are making the decision to take another card in blackjack based on your total and a dealer's up card or you are choosing which route to take home based on likely traffic from past experience, probability is a part of our daily lives. We unknowingly use probability to make routine decisions, but a better understanding of the concepts could actually help in your making the best decisions.

Probability is NOT the same as betting odds (payoffs); Kentucky Derby and Pro Sports sets odds to insure they make money and the odds depend on how the betting goes. Lower odds does not mean more likely to win, just that more people think it is true. Vegas tables set payoffs lower than true odds so they always come out ahead.

Mathematically, probability equals the number of events meeting the specified condition divided by the number of possibilities. With a deck of playing cards, there are 52 possible events. Thus, the probability of drawing an ace from a deck of cards = the number of cards meeting the condition of being an ace (i.e., 4) divided by the number of possibilities (i.e., 52) = $4 / 52$.

An event is just an uncertain outcome, the result of an experiment; an experiment is the process of making an observation. If you flip a coin (experiment), there are two possible events → heads or tails. If you flip a coin several times, you expect 1/2 Heads, 1/2 Tails. If you flip it 10 times you expect 5 Heads, 5 Tails, but what if you get 10 Tails in a row? Has the probability changed? Not really. The probability is still 50% of getting Heads or Tails because what happened in the past has nothing to do with the future (when dealing with an independent event like a coin flip).

Probability is always between 0 and 1 (0% - 100%) inclusive. It is NEVER less than 0% and NEVER greater than 100%.

Solving Probability Problems

Probability problems can easily be solved with tables. Setting up the table is usually the tricky part, but once it is done, the rest is easy. With a deck of cards, you would make a table with the values serving as the columns (A, 2, 3, ..., J, Q, K) and the suits serving as the rows (diamonds, hearts, clubs, spades). Basically, we just employ rules of Set Theory in which you circle the objects in question and count what is in the relevant circle.

Suppose 100 people were asked about their political affiliation and the following shows their responses:

	Dem	Rep	Ind	Total
Male	10	25	5	40
Female	30	20	10	60
Total	40	45	15	100

Now if you were asked the following, you can use simple rules to solve them.

1. Prob[Female]
2. Prob[Republican]
3. Prob[Female and Democrat]
4. Prob[Female or Democrat]
5. Prob[Female / Democrat] (i.e., prob of female given Democrat)
6. Prob[Democrat / Female] (i.e., prob of Democrat given female)

Let's take them one at a time.

1. Prob[Female] = total females / total people = $60/100 = 60\%$

	Dem	Rep	Ind	Total
Male	10	25	5	40
Female	30	20	10	60
Total	40	45	15	100

2. Prob[Republican] = total republicans / total people = $45/100 = 45\%$

	Dem	Rep	Ind	Total
Male	10	25	5	40
Female	30	20	10	60
Total	40	45	15	100

3. Prob[Female **and** Dem] = intersection of females and democrats / total people = $30/100 = 30\%$

	Dem	Rep	Ind	Total
Male	10	25	5	40
Female	30	20	10	60
Total	40	45	15	100

4. Prob[Female **or** Dem] = union of females and dem / total people = $(10+30+20+10)/100 = 70/100 = 70\%$

	Dem	Rep	Ind	Total
Male	10	25	5	40
Female	30	20	10	60
Total	40	45	15	100

5. Prob[Female/Dem] = percentage of democrats who are female (i.e., prob of female given Dem) = # of female democrats / total democrats = $30/40 = 75\%$

	Dem	
Male	10	
Female	30	
Total	40	

6. Prob[Dem/Female] = percentage of females who are democrat (i.e., prob of Dem given female) = # of female democrats / total females = $30/60 = 50\%$

	Dem	Rep	Ind	Total
Female	30	20	10	60

The screenshot on the next page from **Joint_Probability.xls** shows the results of entering the discussed political affiliation data.

Probabilities Calculations		Event B				Use the YEL probability table up to 3 values. If only 2 values are given, use the third column.			
		Dem	Rep	Indep	Totals				
Event A	Male	10	25	5	40				
	Female	30	20	10	60				
	Totals	40	45	15	100				
Simple Probabilities						Joint Probabilities			
Prob(Male)	40%	< --- this is from question 1				Prob(Male and Dem)	10%		
Prob(Female)	60%					Prob(Male and Rep)	25%		
Prob()	0%	this is from question 3 --->				Prob(Male and Indep)	5%		
Prob(Dem)	40%					Prob(Female and Dem)	30%		
Prob(Rep)	45%	< --- this is from question 2				Prob(Female and Rep)	20%		
Prob(Indep)	15%					Prob(Female and Indep)	10%		
Conditional Probabilities						Prob(and Dem)	0%		
Prob(Male given Dem)	25%	T R A V I S K				Prob(and Rep)	0%		
Prob(Male given Rep)	56%					Prob(and Indep)	0%		
Prob(Male given Indep)	33%					Addition Rule			
Prob(Female given Dem)	75%					< --- this is from question 5		Prob(Male or Dem)	70%
Prob(Female given Rep)	44%							Prob(Male or Rep)	60%
Prob(Female given Indep)	67%	this is from question 4 --->				Prob(Male or Indep)	50%		
Prob(given Dem)	0%					Prob(Female or Dem)	70%		
Prob(given Rep)	0%	< --- this is from question 6 -				Prob(Female or Rep)	85%		
Prob(given Indep)	0%					Prob(Female or Indep)	65%		
Prob(Dem given Male)	25%	#DIV/0!				Prob(or Dem)	40%		
Prob(Dem given Female)	50%					Prob(or Rep)	45%		
Prob(Dem given)	#DIV/0!					Prob(or Indep)	15%		
Prob(Rep given Male)	63%								

Excel will solve this for you, but it is important to understand how the answers are derived. Is political party independent of gender? To be independent, $\text{Prob}(A \text{ given } B) = \text{Prob}(A)$. In other words, the probability of an event occurring is unchanged by the existence of the other event being known. The $\text{Prob}[\text{Female}] = 60 / 100 = 60\%$. The $\text{Prob}[\text{Female given Democrat}] = 30 / 40 = 75\%$. Knowing the person's political party alters the probability in regards to gender, so the two variables are dependent.

An example of independent variables would be the suit and value of a playing card. The probability of drawing an Ace from a deck of cards is $4 / 52$. Given that the card drawn is known to be a diamond, the probability that the card is an Ace is now $1 / 13$. Since $4 / 52 = 1 / 13$, the two variables must be independent. Thus, the probability of a card being a specific value is independent of the suit (i.e., knowing the suit will not help you), AND the probability of a card being a specific suit is independent of its value (it works in both directions).

The Normal Distribution

If you were to look at a 5'8" man, you'd probably say he was short. If you were to look at a 5'7" woman, you'd probably say she was tall, despite the fact that she's shorter than the man. Why? It is because we tend to compare each data point to its respective mean. If we wanted to solve problems regarding heights, you would think we would need two separate formulas (one for men and one for women), but it is not true. Thanks to the Normal Probability Distribution, we can put everything on the same scale. The Normal curve is a bell-shaped curve which peaks in the middle at the mean. Units on the curve are measured in terms of the standard deviation; one standard deviation in both directions from the mean captures 68% of the data, two standard deviations in both directions captures 95% of the data, and three standard deviations in both directions captures 99.7% of the data. As expected, the bulk of the data is close to the mean.

Chapter Two: Probability & the Normal Distribution

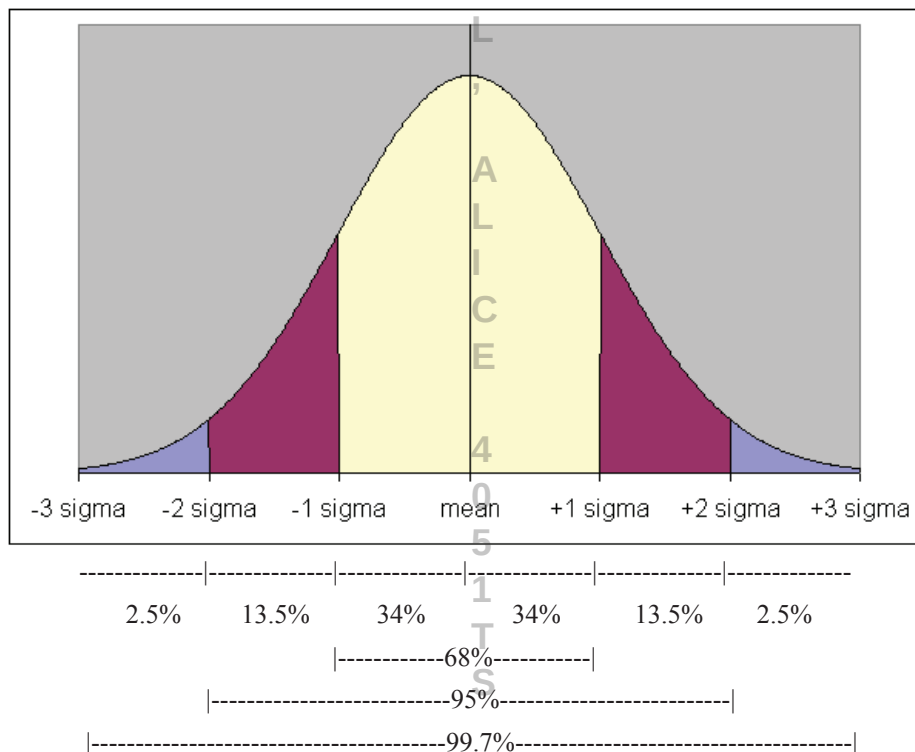
The Standard Normal Distribution has values mainly between -3 and +3 as measured by a z-score, with the z-score being the number of standard deviations a value is from the mean. The formula is

$$z = \frac{x - \mu}{\sigma}$$

Thus the male height of 5'8" is about one standard deviation below average ($z = -1$) and the female height of 5'7" is about two standard deviations above average ($z = +2$). The female would show up to the right of the male on this standard scale.

The Standard Normal Distribution is commonly used. You often hear about standardized testing, but this is what it means. The SAT scores, for example, are computed by comparing one's raw score with the mean raw score and dividing by the standard deviation (thus giving the z-score); then the z-score is converted to the SAT scale, which forces the mean to be 500 and the standard deviation to be 100. If the mean raw score on the math section were 40 and the standard deviation was 10, then a person who scored 60 would have scored 20 points higher than the mean, and since the standard deviation is 10, that person scored 2 standard deviations above average ($z = 60 - 40 / 10 = +2.00$). On the SAT system, this translates to 200 points above the average of 500, so the SAT score would be 700. Since it is normally distributed, we know that about 68% of students score between 400 and 600, 95% score between 300 and 700, and the remaining 5% score below 300 or above 700.

One of the great features of using the Normal Distribution is that we can compute probabilities very easily. We only need the mean and standard deviation to completely define the curve. If you know that you scored 400, and you know the mean = 500 and the standard deviation = 100, you can compute that your score is 1 standard deviation below the mean.

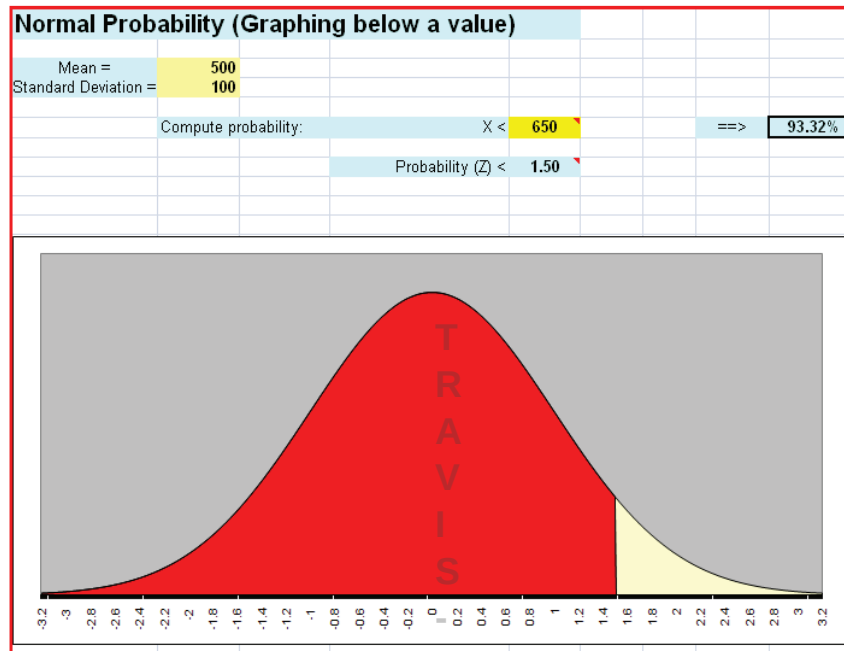


Chapter Two: Probability & the Normal Distribution

If you have ever asked a teacher to grade on a curve, be careful. A true grading curve means that any score more than 1 standard deviation about the mean gets an A, less than one standard deviation about the mean gets a B, up to one standard deviation below the mean gets a C, one to two standard deviations below the mean gets a D, and more than two standard deviations below the mean gets an F. This results in approximately 16% A's, 34% B's, 34% C's, 13.5% D's, and 2.5% F's. Some teachers adjust the curve a bit so that the cutoff for a B is even higher so there are more C's and fewer B's, but the concept is the same. And this means one's grade depends on the grades of others. While a 90% is an A or A- in most schools, if the class mean were 95%, then this score is below average, and depending on the standard deviation, a 90% could be in the C or D range (although we only seem to recall the situations where our poor scores got raised). Standardizing grades mostly benefits those who performed worst (and rarely those who did very well), and the irony is that the curve killer in a class is not the top student but rather the bottom one. A student scoring 100% no doubt raises the class average a bit, but a student scoring 0% will drop the average a bit more while raising the standard deviation an astronomical amount. So in a class of 20 students where the average is 80% and the standard deviation is 10%, the cutoff for an A would be 90%, a B would be 80%, a C would be 70% and a D would be 60%. A grade of 100 added to the mix would likely raise the mean by about one point to 81% while barely impacting the standard deviation; the cutoffs for A, B, C and D might be 91%, 81%, 71% and 61% respectively. But if a grade of 0 were added to the mix instead, the mean would likely drop to around 76% while the standard deviation might increase to over 20%; you might then see a cutoff of 102% for an A, 76% for a B, 50% for a C and 24% for a D. Thus an A would be impossible, a B would be lowered a bit, a C would be lowered a lot, and a student with a solid F can now pass the course, courtesy of a poor student who deprived the A students from getting what they deserve. Rather unfair, but such is the result of standardized scores for small samples (but SATs and other standardized tests are given to over a million students, making this a non-issue).

Normal Probability				
Mean =	500			
Standard Deviation =	100			
Compute probability:	X <	300	==>	2.28%
Compute probability:	X >	700	==>	2.28%
Compute probability:	Range (i.e., between 2 values)	250	<= X <= 450	==> 30.23%
Find actual score:	Percentage ABOVE = (i.e. % greater than)	5.00%	==>	664.49
Find actual score:	Percentage BELOW = (i.e. % less than)	95.00%	==>	664.49

In the **Normal_Probability.xls** file, you can do multiple computations at once, given just a mean and a standard deviation. For example, we know the mean SAT math score is 500 and the standard deviation is 100. We can then compute the probability of scoring above a number, below a number, or between two numbers (this can also be read as the percent of people who scored in those ranges). We can also do the reverse and determine what score corresponds to a certain percentile (in the case above, to score in the top 5% and be at the 95th percentile, you would need to score at least 664.49).



The [Normal_Probability.xls](#) file also includes tabs that show each of the computations graphically. Here we see a computation for the probability of scoring less than 650 on the SAT Math. This corresponds to a z-score of 1.50 ($650 - 500 = 150$, 150 divided by $100 = 1.50$), as 650 is 1.50 standard deviations above the mean. The area under the standard normal curve = 1 or 100%, and the area shaded in red corresponds to the area less than a z of 1.50. The red area accounts for 93.32% of the total area, which means that 93.32% of those taking the SAT score less than 650 on the Math section.

The Central Limit Theorem

When a distribution is not Normal, we cannot compute probability as is. However, if we take samples from the data, the means of the samples will appear normal if the sample size is large enough. Regardless of the look of the original distribution, larger samples will result in the curve starting to appear Normal. As a general rule, a sample of at least 30 guarantees that the distribution of sample means will be normally distributed. The mean of the new distribution is the same as the original mean, but the spread is cut down dramatically and is known as the standard error (which is computed from the standard deviation and the sample size). So instead of solving problems involving individual values, we solve problems involving means, but essentially they are similar.

When solving problems regarding the probability of a sample mean being in a specified range, you only need the mean, the standard deviation AND the sample size. So in the case of the SAT, you might want to know what the probability is of taking a sample of 4 students and getting a mean greater than 600.

Normal Probability for Means (Central Limit Theorem)									
Mean =	500								
Standard Deviation =	100								
Sample Size =	4								
Compute probability:		X-bar <	450			==>		15.87%	
Compute probability:		X-bar >	600			==>		2.28%	
Compute probability:		Range (i.e., between 2 values)	350	<= X-bar <=	450	==>		15.73%	
Find actual mean:		Percentage ABOVE = (i.e. % greater than)	5.00%			==>		582.24	
Find actual mean:		Percentage BELOW = (i.e. % less than)	95.00%			==>		582.24	

Here we see in the Normal-CLT tab of the **Normal_Probability.xls** file that 2.28% of the time, a sample of 4 students will have a sample mean greater than 600. This doesn't mean that 2.28% of the students score above 600. If we used a sample greater than 4, the probability would get smaller because as you take a larger and larger sample, there is a greater likelihood that you will be closer to the true mean. With scores ranging from 200 to 800 on the exam and only 16% scoring above 600, it is hard to imagine taking a large sample and having the mean be greater than 600.

The Central Limit Theorem is quite powerful. With a measure of central tendency and a measure of dispersion, we can completely define a distribution and solve any probability problems; the only catch is that the data must be normally distributed, but if it isn't, there are ways to overcome this and make the data work for you. As we will see in later units, this theorem will form the basis for hypothesis testing and we will take advantage of working with data, whether or not it is normally distributed.

CHAPTER TWO KNOWLEDGE ASSESSMENT

Probability & the Normal Distribution

Discussion Questions

DISCUSSION QUESTION 1

SAMPLING VS. RELIABILITY: “The larger the sample, the more reliable the results.” Do you agree or disagree with this statement? Explain.

DISCUSSION QUESTION 2

SEEING THROUGH THE CLAIM: An auto manufacturer advertises that “90% of the cars we’ve ever made are still on the road.” Assuming this is literally true, how can it be explained? What facts / statistics would you need to know to expose this misleading claim?

Practice Problems: Real Estate

Solutions are provided to practice problems so you can check your work.

Use the Real_Estate.xls file which consists of 100 homes purchased in 2007 and appraised in 2008. It includes variables regarding the number of bedrooms, number of bathrooms, whether the house has a pool or garage, the age, size and price of the home, what the house is constructed from, and the appraisals from two agents.

PRACTICE PROBLEM 1:

Create a pivot table of Pool and Garage. Then complete the Joint Probability table so you can answer the following:

- What is the probability of randomly choosing a home that has no pool?
- What is the probability of randomly choosing a home that has a pool AND no garage?
- What is the probability of randomly choosing a home that has a pool OR a garage?
- Given that the home you selected has a pool, what is the probability it also has a garage?

PRACTICE PROBLEM 2:

Let’s assume that the Real_Estate.xls file was the entire population. We know the mean and standard deviation of home sizes to be 2212 sq. ft. and 235 sq. ft., respectively. Using the Normal_Probability.xls file, compute the percentage of homes that are

- smaller than 2500 sq. ft.?
- larger than 2300 sq. ft.?
- between 2000 and 2100 sq. ft.?
- The largest 20% of homes are greater than what size?

In each case, compare the computed results to the truth as found in the actual data file.

CHAPTER TWO KNOWLEDGE ASSESSMENT

Probability & the Normal Distribution

PRACTICE PROBLEM 3:

Knowing the mean and standard deviation of home sizes to be 2212 sq. ft. and 235 sq. ft., respectively, if we were to visit 4 homes at random every day and compute the mean size, what is the probability that the mean would be

- a) smaller than 2500 sq. ft.?
- b) larger than 2300 sq. ft.?
- c) between 2000 and 2100 sq. ft.?
- d) 20% of the time what would you expect the mean to be above?

Assigned Problems: Student Data

Use the Student_Data.xls file which consists of 200 MBA at Whatsamattu U. It includes variables regarding their age, gender, major, GPA, Bachelors GPA, course load, English speaking status, family, weekly hours spent studying.

ASSIGNED PROBLEM 1:

Create a pivot table of Gender and Major. Then complete the Joint Probability table so you can answer the following:

- a) What is the probability of randomly choosing a Female?
- b) What is the probability of randomly choosing a Male AND Finance major?
- c) What is the probability of randomly choosing a Female OR Leadership major?
- d) Given that the home you selected is a Male, what is the probability he has no major?
- e) Given that the home you selected has no major, what is the probability the student is male?

ASSIGNED PROBLEM 2:

Let's assume that the Student_Data.xls file was the entire population. We know the mean and standard deviation of student ages to be 42.3 and 8.9, respectively. Using the Normal_Probability.xls file, compute the percentage of students that are older than 50, younger than 40, between 41 and 46, and oldest 10% are at what age? Then compare to the truth as found in the actual file.

APPENDIX

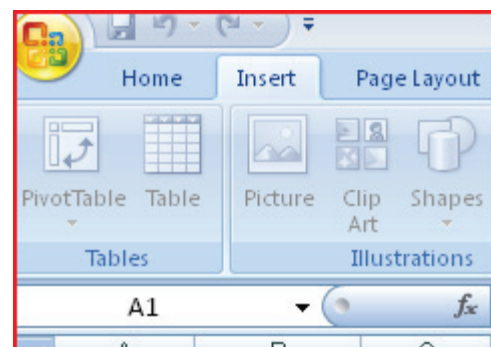
PIVOT TABLE TOOL TUTORIAL

In this tutorial, you will learn to utilize the Pivot Table tool built into Excel. This tool allows you to create crosstabulations and to dig into a data file very deeply, grouping the data as you wish, and even analyzing with statistical options and graphs.

We will use the Real Estate data from the course, which consists of 100 homes. First, open the file and select the data with your mouse (to row 101). It is imperative that the variable names be in the first row.

	A	B	C	D	E	F	G	H	I	J	K
1	House_No	Bedrooms	Size	Age	Pool	Construction	Garage	Baths	SalePrice	Appraisal_1	Appraisal_2
2	1	4	2,349	8	0	Brick	1	2	263,000	237,000	249,000
3	2	4	2,102	7	1	Wood	0	2	182,000	164,000	156,000
4	3	3	2,271	6	1	Brick	0	2	242,000	219,000	228,000
5	4	2	2,188	6	1	Stucco	0	2.5	214,000	194,000	186,000
6	5	2	2,148	10	1	Stucco	0	1.5	140,000	127,000	131,000
7	6	2	2,117	2	0	Stucco	1	2	245,000	223,000	216,000
8	7	5	2,484	7	1	Brick	1	2	300,000	298,000	304,000
9	8	2	2,130	6	1	Stucco	1	2.5	272,000	249,000	244,000
10	9	3	2,254	2	0	Stucco	0	1.5	221,000	202,000	204,000
11	10	4	2,385	3	1	Wood	1	2	267,000	245,000	243,000
12	11	4	2,108	4	1	Brick	1	2	292,000	269,000	282,000
13	12	2	1,715	5	1	Wood	1	1.5	209,000	193,000	183,000
14	13	5	2,495	1	1	Wood	1	2	271,000	250,000	260,000
15	14	4	2,073	6	1	Brick	1	2	246,000	228,000	219,000
16	15	2	2,283	2	1	Brick	0	2	194,000	180,000	185,000
17	16	3	2,119	7	1	Stucco	1	2	281,000	261,000	253,000
18	17	4	2,189	2	0	Brick	0	2	173,000	161,000	164,000
19	18	5	2,316	1	0	Wood	0	2.5	207,000	193,000	189,000
20	19	3	2,220	5	0	Wood	1	2	199,000	186,000	188,000
21	20	5	1,901	4	0	Wood	1	2	209,000	196,000	194,000
22	21	4	2,624	8	1	Wood	1	2	252,000	237,000	249,000
23	22	4	1,938	4	0	Stucco	1	2.5	193,000	182,000	173,000
24	23	5	2,101	6	1	Brick	0	1.5	209,000	197,000	205,000
25	24	5	2,141	9	1	Brick	1	3	320,000	309,000	318,000

Under the INSERT menu, choose the PIVOT TABLE icon. You will then see an option for PIVOT TABLE and PIVOT CHART – choose the PIVOT TABLE.



The CREATE PIVOT TABLE box pops up with the range already entered (if you highlighted the data first). It is also defaulted to create a new worksheet. Click OK.

	A	B	C	D	E	F	G	H	I	J	K
1	House_No	Bedrooms	Size	Age	Pool	Construction	Garage	Baths	SalePrice	Appraisal_1	Appraisal_2
2	1	4	2,349	8	0	Brick	1	2	263,000	237,000	249,000
3	2	4	2,102	7	1	Wood	0	2	182,000	164,000	156,000
4	3	3	2,271	6	1	Brick	0	2	242,000	219,000	228,000
5	4	2	2,188	6	1	Stucco	0	2.5	214,000	194,000	186,000
6	5	2	2,148	10	1	Stucco	0	1.5	140,000	127,000	131,000
7	6	2	2,117	2	0						
8	7	5	2,484	7	1						
9	8	2	2,130	6	1						
10	9	3	2,254	2	0						
11	10	4	2,385	3	1						
12	11	4	2,108	4	1						
13	12	2	1,715	5	1						
14	13	5	2,495	1	1						
15	14	4	2,073	6	1						
16	15	2	2,283	2	1						
17	16	3	2,119	7	1						
18	17	4	2,189	2	0						
19	18	5	2,316	1	0						
20	19	3	2,220	5	0						
21	20	5	1,901	4	0						
22	21	4	2,624	8	1						
23	22	4	1,938	4	0						
24	23	5	2,101	6	1	Brick	0	1.5	209,000	197,000	205,000

Create PivotTable

Choose the data that you want to analyze

☒ Select a table or range

Table/Range: RealEstate!\$A\$1:\$K\$101

☐ Use an external data source

Choose Connection...

Connection name:

Choose where you want the PivotTable report to be placed

☒ New Worksheet

☐ Existing Worksheet

Location:

OK Cancel

PivotTable Field List

Choose fields to add to report:

☐ House_No

☐ Bedrooms

☐ Size

☐ Age

☐ Pool

☐ Construction

☐ Garage

☐ Baths

☐ SalePrice

☐ Appraisal_1

☐ Appraisal_2

Drag fields between areas below:

Report Filter

Column Labels

Row Labels

Values

You will now see a grid on the left where you can drop the row, column and data items. On the right, you will see a PIVOT TABLE FIELD LIST based on the variables in your data file.

PivotTable Field List

Choose fields to add to report:

☐ House_No

☐ Bedrooms

☒ Size

☐ Age

☐ Pool

☒ Construction

☐ Garage

☐ Baths

☐ SalePrice

☐ Appraisal_1

☐ Appraisal_2

Drag fields between areas below:

Report Filter

Column Labels

Row Labels

Values

Construction

Sum of Size

Click on CONSTRUCTION and it will automatically populate the ROW LABELS box. Now click on SIZE and while holding the mouse button down, drag the label into the VALUES box.

You will now see that it shows SUM OF SIZE. This means that it adds up the sizes, which we do not want. This is a default with numerical data and is easily changed. Click on the arrow next to SUM OF SIZE and then choose VALUE FIELD SETTINGS. This will open the box below.

Count of Size	
Construction	Total
Brick	40
Stucco	35
Wood	25
Grand Total	100

Change to COUNT and click OK.

Value Field Settings

Source Name: Size
Custom Name: Sum of Size

Summarize by Show values as

Summarize value field by
Choose the type of calculation that you want to use to summarize the data from selected field
Sum
Count
Average
Max
Min
Product

Number Format OK Cancel

Truthfully it doesn't matter what variable you used for values when you are just doing a COUNT as long as it is a variable that doesn't have missing values. You now see a table showing that of the 100 homes, 40 were made of Brick, 35 of Stucco and 25 of Wood.

Here's something interesting that you should see and can use at any time in any pivot table. Double click on cell B7 which shows a 25 in it. Instantaneously you get a new spreadsheet showing all variables for those 25 Wood homes. It is on its own worksheet, so you can save it separately if you wish. For now let's go back to the Pivot Table.

	A	B	C	D	E	F	G	H	I	J	K
1	House_No	Bedrooms	Size	Age	Pool	Construction	Garage	Baths	SalePrice	Appraisal_1	Appraisal_2
2	99	4	2132	3	1	Wood	1	2	174000	192000	184000
3	2	4	2102	7	1	Wood	0	2	182000	164000	156000
4	96	3	2442	2	1	Wood	1	2	179000	197000	207000
5	92	4	2886	10	1	Wood	1	2	227000	247000	240000
6	91	5	2091	1	0	Wood	1	2	188000	205000	211000
7	90	3	2442	6	0	Wood	1	2	179000	194000	186000
8	89	5	2548	1	0	Wood	0	2.5	187000	203000	211000
9	84	2	1887	4	1	Wood	1	1.5	188000	202000	206000
10	82	4	2624	2	1	Wood	1	2	240000	256000	264000
11	10	4	2385	3	1	Wood	1	2	267000	245000	243000
12	74	4	2312	3	1	Wood	0	2	164000	173000	176000
13	12	2	1715	5	1	Wood	1	1.5	209000	193000	183000
14	13	5	2495	1	1	Wood	1	2	271000	250000	260000
15	72	2	1871	8	1	Wood	0	1.5	125000	131000	135000
16	61	5	2170	4	0	Wood	1	2.5	270000	276000	287000
17	48	3	2014	8	0	Wood	0	2	172000	171000	173000
18	44	4	2262	5	0	Wood	1	2	176000	174000	179000
19	18	5	2316	1	0	Wood	0	2.5	207000	193000	189000
20	19	3	2220	5	0	Wood	1	2	199000	186000	188000
21	20	5	1901	4	0	Wood	1	2	209000	196000	194000
22	21	4	2624	8	1	Wood	1	2	252000	237000	249000
23	43	3	2012	3	0	Wood	0	2	205000	202000	194000
24	35	4	2023	5	0	Wood	0	2.5	183000	178000	173000
25	27	2	2117	7	1	Wood	1	2	257000	245000	240000
26	26	2	1912	5	1	Wood	0	2	187000	178000	182000

Click on the arrow next to COUNT OF SIZE and then choose VALUE FIELD SETTINGS. This time when the box pops up, go to the tab labeled SHOW VALUES AS.

Value Field Settings

Source Name: Size

Custom Name: Count of Size

Summarize by Show values as

Show values as

% of total

Base field: House No, Bedrooms, Size, Age, Pool, Construction

Base item:

Number Format OK Cancel

Note that the Pivot Table now shows the totals in percents. So we see that 40% of the homes are made of Brick, 35% Stucco and 25% Wood.

Count of Size	
Construction	Total
Brick	40.00%
Stucco	35.00%
Wood	25.00%
Grand Total	100.00%

Let's go back to the VALUE FIELD SETTINGS but this time show the data as NORMAL again and change from a COUNT to an AVERAGE. Then click OK.

Value Field Settings

Source Name: Size

Custom Name: Count of Size

Summarize by Show values as

Show values as

Normal

Base field: House No, Bedrooms, Size, Age, Pool, Construction

Base item:

Number Format OK Cancel

Value Field Settings

Source Name: Size

Custom Name: Average of Size

Summarize by Show values as

Summarize value field by

Choose the type of calculation that you want to use to summarize the data from selected field

Sum, Count, Average, Max, Min, Product

Number Format OK Cancel

Average of Size	
Construction	Total
Brick	2239.65
Stucco	2173.171429
Wood	2220.12
Grand Total	2211.5

Now you can see the average Size for the different constructions. A Brick home has an average size of 2239.65 square feet, for instance.

Count of Size	
Construction	Total
Brick	40
Stucco	35
Wood	25
Grand Total	100

Now let's reset it back to a COUNT using the VALUE FIELD SETTINGS.

PivotTable Field List

Choose fields to add to report:

- ☐ House_No
- ☐ Bedrooms
- ☒ Size
- ☐ Age
- ☒ Pool
- ☒ Construction
- ☐ Garage
- ☐ Baths
- ☐ SalePrice
- ☐ Appraisal_1
- ☐ Appraisal_2

Drag fields between areas below:

Report Filter:

Column Labels: Pool

Row Labels: Construction

Values: Count of Size

Let's add another variable to the table. Drag POOL into the COLUMN LABELS.

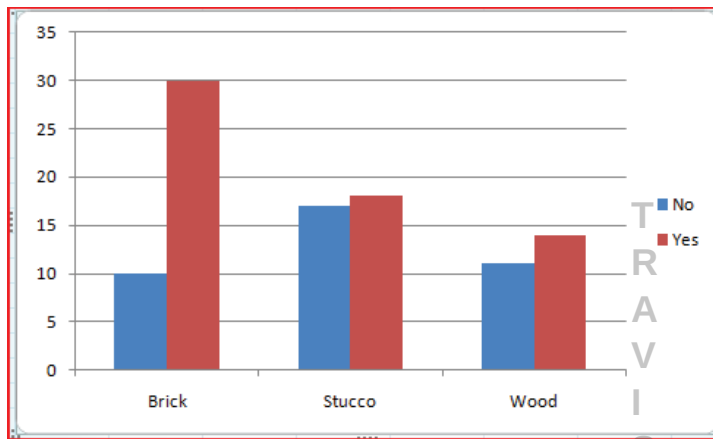
We now have a two-dimensional crosstabulation. Note that 0 means NO POOL and 1 means POOL. You can actually type right over the 0 and 1 to change these labels (as shown below).

Count of Size	Pool		
Construction	0	1	Grand Total
Brick	10	30	40
Stucco	17	18	35
Wood	11	14	25
Grand Total	38	62	100

Count of Size	Pool		
Construction	No	Yes	Grand Total
Brick	10	30	40
Stucco	17	18	35
Wood	11	14	25
Grand Total	38	62	100

Here we see that 10 of the Brick homes did not have a pool while 30 did. And the Stucco homes were almost evenly split on having or not having a pool. At this point you can go back to displaying percentages if you wish, and even choose to show the percents by row, by column or by the overall total (e.g., if shown by row, you would see 25% of the Brick homes did not have a pool and 75% did).

Now click on any of the numbers in the Pivot Table. Then click on the COLUMN option on the toolbar (under INSERT in the CHARTS section). For now just choose the chart in the top row on the left side (under 2-D Charts).



You can instantly convert your Pivot Table into a useful graph. If you wish to edit it further, you need only click on the graph. For now, click on the outer frame of the chart and hit the DELETE key.

While we needn't do it here, click on the arrow next to CONSTRUCTION. Notice the check boxes. If you unchecked Stucco, you would then only be displaying the Brick and Wood homes. And if you remove the CONSTRUCTION variable from the Pivot Table and put it back later, Stucco would still be omitted until you restore it.

Now for one last but important tool / grouping. Often we have a situation where we want to recode data into groups for various reasons and rather than altering the data file, we can do it so easily in a Pivot Table. A common example is to take survey responses and recode (group) Strongly Agree and Agree responses into a category titled POSITIVE, and other responses into a category titled NEGATIVE. Since our data set doesn't have such variables, let's use the SALES PRICE.

Under COLUMN LABELS and ROW LABELS (leave the COUNT OF SIZE alone), click on the arrows next to each line and choose REMOVE FIELD. This will blank out your Pivot Table. You could have also grabbed the variable from the table and dragged it out.

Count of Size	Total
Total	100

Count of Size	
SalePrice	Total
125,000	1
126,000	1
140,000	1
147,000	1
154,000	1
155,000	1
164,000	1
166,000	2
172,000	2
173,000	2
174,000	2
175,000	1
176,000	3
177,000	1
179,000	2
180,000	1
182,000	1
183,000	1
187,000	2
188,000	5
189,000	1
191,000	1
192,000	1

Now click on SALEPRICE and drag it into the ROW LABELS.

There are 74 different prices in the 100 homes. Such a display is not very helpful since it is mostly 1's with a few 2's. But we can turn this scale data into an ordinal variable by putting the data into ranges.

T
R
A
V
I
S
-
K
E
N
D
A
L
L
A
I
C
E
4
0
5
1
T
S

Count of Size		
SalePrice2	SalePrice	Total
Group1	125,000	1
	126,000	1
	140,000	1
	147,000	1
	154,000	1
	155,000	1
	164,000	1
	166,000	2
	172,000	2
	173,000	2
	174,000	2
	175,000	1
	176,000	3
	177,000	1
	179,000	2
	180,000	1
	182,000	1
	183,000	1
	187,000	2
	188,000	5
	189,000	1
	191,000	1
	192,000	1
	193,000	2
	194,000	1
	198,000	1
	199,000	2

First highlight cells A5 through A31, capturing the 125000 to the 199000. Then right click your mouse and choose the GROUP option.

A new variable was created called (SALEPRICE2) with a value of Group1 under it. If you click on the Group1, you can rename it (change it to <200k).

Count of Size		
SalePrice2	SalePrice	Total
< 200k	125,000	1
	126,000	1
	140,000	1
	147,000	1
	154,000	1
	155,000	1
	164,000	1
	166,000	2

Now the 200000 through the 248000 and click on the GROUP option again. And repeat this for the rest of the data, from 251000 to 320000. Now rename Group2 to 200-250k, and rename Group3 to >250k.

And then drag SALEPRICE out of the table (or click on the arrow next to the variable name and choose REMOVE FIELD). You now have a table with a new variable, which actually appears in the variable list on the right of the screen. From the table you can see that 41 of the homes sold for less than \$200,000 and 23 sold for over \$250,000.

Count of Size	
SalePrice2	Total
< 200k	41
200-250k	36
>250k	23
Grand Total	100

We can go on further, as this is not the depth of Pivot Tables, but it covers the basics and even some advanced stuff. By playing with it further, you will likely discover even more, but most of what you would do with this wonderful tool has been addressed here. With little effort you will see that this tool is fun to use and quite amazing.

T
R
A
V
I
S
-
K
E
N
D
A
L
L
,
A
L
I
C
E

4
0
5
1
T
S

CHAPTER THREE

CONFIDENCE INTERVALS & SAMPLE SIZES

Sampling

Statistics means never having to say you're certain. Or at least that how it feels since we discuss everything in terms of probability and error. Remember that statistics are all about using samples to infer the truth about populations. When we take a sample, we would love to believe that whatever is true about the sample is true about the population, but that will rarely ever happen, as there is almost always some difference (known as a margin of error or sampling error). The better the sample is, the smaller the margin of error should be, and so understanding the nature of sampling is critical to understanding statistics.

Let's start with why we sample at all. Wouldn't you rather have all of the data in the population and thereby have no sampling error? Of course, but we choose not to for several reasons.

1. **Cost** – it costs money to purchase mailing lists or phone lists or even email lists. It costs even more to hire people to mail / phone / email the surveys, and going door-to-door to survey everyone is obscenely expensive. Add to that the fact that phone calls are not free, and mailing surveys involves postage for the survey itself and the return envelope. Then factor in the low return rate on surveys, meaning that you would have to keep calling the same people or mailing out second and third rounds of surveys. Clearly not an inexpensive venture, which is part of the reason we only conduct a U.S. census once every ten years.
2. **Time** – To gather 400 responses for a phone survey, a marketing research firm will often call over 4,000 people, factoring the nonresponses they will likely get. And calling people multiple times in the hopes of reaching them adds more time to the process. It could take a week for a typical sized firm to conduct such a poll. Now imagine the time involved if they had to survey the population, and then proceed to analyze all of that data which must be loaded into a computer. Our census takes most of the year to conduct and they have a lot of people working full time on nothing but the census. So even if money is no object, time is still a major consideration when polling a large population.

3. **Accuracy** – While it may seem to defy logic, samples are often more accurate than populations. With a sample, we accept that it is not perfect and there is a built-in margin of error. And if we choose a fairly sized representative sample, the error is minimized. Yet, if we aim to get the entire population, we are expecting perfection. Do you truly believe that the U.S. Census has the exact average age in the U.S.A., for example? That not one person failed to receive a form, that everyone completed their forms, that none of the forms got lost in the mail en route to the Census Bureau, and that all the data was entered in a database without error? If even one form is missing or one number is mistyped, the population parameters are wrong. But in entering a sample of 1,000, for example, we can easily triple-check our work and even contact the respondent to make sure the data is accurately entered. It is more preferable to report a statistic with a small margin of error we feel confident in than to report a parameter with no alleged error that we have little confidence in. So unless collecting the population data involves little more than downloading the data from the web (e.g., the population of NYSE stock prices on a given day), a well-chosen sample is often more preferable for the sake of accuracy.
4. **Availability of population** – Whether or not time and money are on your side doesn't matter if the population is not available. While some populations may be a captive audience (e.g., all students attending a grade school – can be surveyed in school over several days to capture any absentees), others are not quite as easy to reach. Do you really believe the Census captures everyone, including all of the homeless, all of those living in the mountains away from civilization, all of those visiting abroad, all of those hospitalized, and all of those in the military amidst a war zone? Let's not forget those who are home but refuse to answer their phones or open their mail. If you cannot reach everyone in the population, you cannot possibly compute the parameter, and if you mistakenly think you captured the population, you are misrepresenting your results.
5. **Destructiveness of observations** – While this is not an issue with simple surveys, some data collection involves the destruction or consumption of the item being studied. To test the life of all batteries produced by Duracell would mean running them until they die and recording the lifespan, and then having nothing left to sell. To taste test cookies would mean taking a bite and packaging the partially chewed cookie for sale (not very appetizing for the customer). For many products, it is not practical or advisable to test an entire population, even at the risk of selling a customer a defective product.

Now that it is apparent why we choose to sample, it is important that the sample be a good one or else the data is meaningless. Selecting a sample involves six stages.

1. **Define the target population** – Before proceeding to get any data, you must define the group that the data is representing. An election poll would be meaningless if the people surveyed were nonvoters or those from the wrong voting district or even those who are eligible but not registered voters. As an example, if a study were being done at a company to see how satisfied employees were with the performance review process, the target population in question would be only those employees who participated in the process in

the past year (excluding newer employees and those who haven't been involved in a recent performance review for whatever reason). Once we know the population we are targeting, we can then figure out how best to represent them without surveying all of them.

2. **Select a sampling frame** – A sampling frame is merely the source from which we get our sample. This could include a physical list, such as a phone book or mailing list, or it could be more of a situation, such as stating that you will select every 10th person who shows up in line (i.e., the sampling frame is being in line). With every sampling frame, there is likely to be some problems, such as missing people or people that should not be included. Judging a sampling frame comes down to asking whether it really misrepresents the truth. Is a residential phone book a poor choice of a sampling frame... it depends! And it is not because not everyone is in it. If you are using a phone book to ask people about their professions, then it is poor since certain professions, such as law enforcement and doctors and judges, are often unlisted and the results will be misleading. Yet if you use the same phone book to survey people about whether they prefer smooth or crunchy peanut butter, the question becomes whether you think that the results of the survey would be different if unlisted people were included (like do unlisted people differ significantly from listed people in their preference for peanut butter?). Since it is hardly likely to bias or results, we would happily use the phone book and assume the results of the survey reflect the entire population.
3. **Choose a probability or non-probability sampling method** – Samples come in two major categories, random and non-random (otherwise known as probability and non-probability). A random sample is essentially one in which all members of a population have an equal chance of selection and is representative of the population, while a non-random sample is selected without regard to their probability of occurrence and is only representative of the sample itself. Non-random samples come in many forms and have the advantage of being cheaper and faster than random ones, but are not generalizeable to the population, lack objectivity and introduce more opportunities for bias:
 - **Convenience sample** – this is essentially a mall survey in which people are chosen out of convenience because they are there. Images of people carrying clipboards by the mall escalator probably come to mind. They don't care who they survey or that it is a representative sample; they merely want warm bodies to complete surveys. While not typically the best option for a survey, they are often a useful option. If you wanted to know what customers think of a new line of coffee at your restaurant, the most practical way to get this data is to offer samples to those in line, but you don't care if it is representative of the population (you merely want to know if it is worth selling in your store).
 - **Judgment sample** – this is a sample in which experts choose the members. Consider the Dow Jones Industrial Average in which 30 stocks are chosen to represent the New York Stock Exchange (NYSE), or the Consumer Price Index (CPI) in which a few select products are chosen to determine the rate of inflation. Clearly not an ideal sample, but

because of the expertise used, the sample size can be kept small and the value of the sample is regularly monitored (note: a mere 30 stocks does a fair job at gauging the entire NYSE, and it has been in use for over a century).

- **Snowball sample** – this is a sample in which it literally grows by snowballing. Think of all of the political surveys sent out over the web; if the respondent feels it is important (like asking if you would favor a tax increase), that person would likely forward the link to many friends to also participate, and they would do likewise. The result would be that more people respond to it than were actually solicited to respond. And while we like to see large sample sizes, since the person conducting the survey has no way of knowing who participated, it would be inappropriate to try to draw inferences to the population, and the statistics are not as meaningful as they would have been had the original sample responded and kept it to themselves.
- **Voluntary sample** – this is a sample in which respondents throw themselves into the study without an invitation. When a poll is on a website for anyone to answer, there is no control in place as to who participates or how often, and the results are questionable. Such is the case with the American Idol voting too, as they are unsolicited votes from a voluntary sample, often with repeated votes (but the folks at AI don't really care about the best person winning or the one that most viewers want, but that they get a lot of audience participation since they are in the business of making money and selling ads).

Random samples can also take many forms but they all have a similar goal in mind of representing the population. Compared to the non-random samples, they are more costly and more time-consuming, but minimize bias and have more generalizeable results.

- **Simple random sample** – this is essentially Lotto, in which all members of the population are tossed in the mix and a sample is chosen, one at a time. Truly a fair sample with everyone having an equal chance at selection, but it can be slow to choose one at a time (unless it is automated on a computer). There are also some inherent risks of leaving out a subgroup by random chance (e.g., randomly choosing 10 people from a class and somehow choosing all men or all women even though there are 50 of each in the room – can certainly happen, even though it is less likely with larger samples).
- **Stratified random sample** – To combat the problem of accidentally not representing a subgroup, you can stratify the data into subgroups and then choose randomly from each. So if a class consisted of 60 men and 40 women, and you wanted to sample 10 people, you would randomly choose 6 of the 60 men, and separately choose 4 of the 40 women. As a result you would have a sample that mirrored the population proportionally and wouldn't have problems with a group being omitted. You can even subgroup further, like making sure to get a representative group of Male freshmen, Female freshmen, Male sophomores, etc.

- **Systematic random sample** – Instead of pulling one item at a time from a group, a far simpler approach to getting the data randomly is to take every n th item. For example, if you wanted to sample 50 from a group of 100, you could flip a coin and either take the 1st, 3rd, 5th, etc. or the 2nd, 4th, 6th, etc. If you only wanted 25, you could randomly pick one of them and then take every 4th thereafter. This can be done on the whole group or each subgroup (so you could stratify your data and pick the 6 men from the 60 by choosing every 10th and pick the 4 women from the 40 women by choosing every 10th). Just be careful not to get carried away with too many subgroups or it gets challenging (such as gender and class year and hair color and major all at once – many subgroups may only have 1 or 0 in them and you won't get proportional results but will be forced to choose an entire subgroup).
 - **Cluster random sample** – This is truly the quick-and-dirty method of getting a random sample. If the data is in clusters of sorts, you can randomly choose the clusters and then all members of those clusters are included in the sample. So if you wanted to conduct a survey at a university, and wished to survey 1,000 students, instead of mailing out 1,000 surveys and hoping for the best (at great expense) or picking a couple of students per class and disrupting almost every class, you could randomly choose 50 course sections (assuming an average of 20 per class) and then survey the entire class in each case. This is cheaper and faster but there is a risk of misrepresentation if you choose very few clusters (e.g., you might miss out on the freshmen classes or the evening classes easily). You could once again combine methods by choosing clusters within each strata, and choose X of the freshmen courses, Y of the evening courses, etc. and make sure that you get an even better cross-section of the university.
4. **Determine the sample size** – While the methods used to compute the sample size will be addressed later in this unit, it would help to deal with any misconceptions first. Ask yourself whether you would trust the results of a sample of 500 chosen from a population of 10,000 students (5% of the population). Now ask yourself if you would trust the results of a sample of 500 chosen from 1,000,000 in your city (0.05%). And how about a sample of 500 chosen from across the USA's 300,000,000 (0.00017%). Would you be surprised to learn that they are EQUALLY GOOD? That's right...the size of the population means nothing here. If this doesn't make sense, then think about this. You are making a two-gallon pot of five alarm chili for your family and then found out that your grandmother was coming over, so you decided to also make a separate one-quart pot of mild chili. Now you wish to see if they are ready to serve. You take a spoon and stir the mild chili and taste it... mmm good. Now here's the question ... how many spoonfuls of the other chili would you need to taste? There is 8 times as much chili in the larger pot, and yet one spoonful will tell you as much about the larger pot as it will about the smaller pot. And the key here is that you first stirred the pot. As long as the chili is well-mixed, the sample is representative of the population and you don't need a large sample to do the job. Most polls are conducted with samples of 400 – 1,000 and we put a lot of trust in the results.

- 5. Choose a data collection technique** – Data is collected in so many ways, and the method varies with the budget, audience and type of data being collected. We now collect most data by phone, mail, email, in person, on the web, and by text messaging. To help in making the decision as to the best approach, look at the table below which compares and contrasts several techniques.

	Personal	Phone	Mail	E-Mail
Costs	<i>High</i>	<i>Medium</i>	<i>Low</i>	<i>Very Low</i>
Time required	<i>Medium</i>	<i>Low</i>	<i>High</i>	<i>Low</i>
Data quantity per respondent	<i>High</i>	<i>Medium</i>	<i>Low</i>	<i>Medium</i>
Reaches widespread sample	<i>No</i>	<i>Maybe</i>	<i>Yes</i>	<i>Yes</i>
Reaches special locations	<i>Yes</i>	<i>Maybe</i>	<i>No</i>	<i>Yes</i>
Interaction with respondents	<i>Yes</i>	<i>Maybe</i>	<i>No</i>	<i>No</i>
Degree of interviewer bias	<i>High</i>	<i>Medium</i>	<i>Low</i>	<i>Low</i>
Severity of non-response bias	<i>Low</i>	<i>Low</i>	<i>High</i>	<i>High</i>
Presentation of visual stimuli	<i>Yes</i>	<i>No</i>	<i>Maybe</i>	<i>Yes</i>

- 6. Select your sample** -- The stage is set, so now you can choose your sample to minimize error and maximize representativeness.

Confidence Intervals

Sampling is the essence of statistics. We normally cannot get access to an entire population and must make do with a sample, but we know that a sample is not exactly the same as the population. There is going to be some difference between the results of a sample and the truth, and if we took another sample, the results would be different again. Ideally, we would like for the results of the sample to be relatively close to the truth, and for us to have confidence in the results. Remember that the real objective in statistics can best be summarized in the definition of inferential statistics -- using a sample to make inferences about a population.

When we make inferences, we are estimating. We rarely know the value of a population mean, but we can compute a sample mean. And no matter how many times you draw a sample, the numbers will vary and the sample mean will not likely equal the population mean exactly. You can randomly stop at 10 gas stations, and compute the mean gas price, but it is doubtful the mean will equal the population mean; yet if you visit gas stations in different locations and with different brands, you will likely come closer to the true mean than if you just picked 10 nearest your home or 10 that were all of the same brand. Considering the sample mean by itself is not accurate and may be too large or too small, we use an interval to estimate. This “confidence” interval is basically the sample mean + / - a margin of error (often called a sampling error since it is the difference between the results from a sample and the actual truth of the population).

To better understand how this works, imagine being in a war zone in 1942 and watching an enemy aircraft fly over and drop a bomb from the cockpit, aiming for some target unknown to you (let's call him 5 o'clock Charlie, based on a classic MASH episode). Once he hits his target, he leaves the area and doesn't return. The bomb hits a Starbucks, but was that the target? Was Charlie the bombardier so perfect with his accuracy that on his first attempt he hit the target? Or did he have poor aim and / or wind conditions which caused him to miss by a lot? Or perhaps he barely missed the target? Without knowing more about Charlie and the conditions, our sample of one bomb leads us to infer that the target was on Starbucks give or take a large margin of error (perhaps a mile in any direction). Suppose he came back at 5 o'clock the next day and dropped a single bomb that hit a Caribou Coffee Shop a half mile East of the Starbucks. Does Charlie have something against coffee? Was this his actual target which he now right on the second try? Or did he overcorrect from the first drop and miss by a bunch in the other direction? Who knows for sure, but if I had to guess, I would estimate the target is mid-way between the two coffee shops give/take a smaller margin of error (perhaps $\frac{3}{4}$ of a mile) based on the fact that his first two bombs were a half mile apart. This pattern would continue – after 10 bombs, I would estimate the target to be in the center of the 10 drops give/take a margin of error based on the spread of the drops. Basically you would estimate the location based on the mean of the sample, and your margin of error would be based on the spread of the sample data, the size of the sample, and how sure you wish to be about capturing the true target. And such is how confidence intervals work.

Here's another illustration that might help further. Imagine if you were asked to throw a dart at the bullseye on a dartboard from 12 feet away. If you were given one throw, you probably wouldn't have much confidence in your ability to hit the target. If the target were expanded to include the ring around the bullseye, your confidence would grow. If the target were expanded further to include the entire dartboard, your confidence would grow more. Essentially, the larger the margin of error, the greater your confidence at hitting the target. Now what if you were aiming for the ring around the bullseye again and you were given several practice throws. If you hit the target 8 out of 10 times and you were asked where you expect the next dart to hit, you would probably draw a small target. The expected target is probably much smaller than it was prior to your first throw. Essentially, the larger the sample size, the smaller the margin of error.

Now for the math. We express a confidence interval by stating that the population mean = the sample mean +/- the sampling error. The sampling error is computed as the z-score (based on the level of confidence) times the standard deviation divided by the square root of the sample size.

$$S.E. = z * S.D. / \sqrt{n}$$

We are essentially saying that we are X% confident that the true population mean is within the confidence interval.

With a 95% confidence interval, we use a z of 1.96 since 95% of the data in a normal distribution is within 1.96 standard deviations of the mean. For a 90% confidence interval we use a z-score of 1.645. For a 99% confidence interval we use 2.576. Thus the higher the level of confidence, the larger the margin of error, and thus the wider the confidence interval.

The sample size is in the denominator of the sampling error, so the larger the sample, the smaller the error and thus the smaller the confidence interval. This makes sense since more data would result in having more information about the population and being closer to the truth.

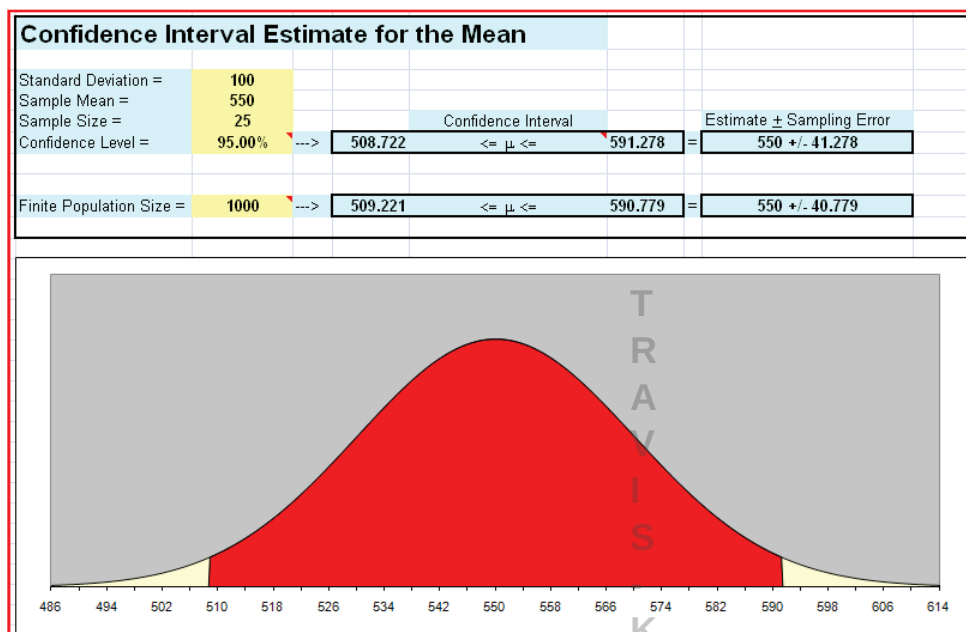
The z-score corresponds to the standard normal distribution. A 95% confidence interval captures the middle 95% of the normal curve, so the z-score is 1.96 since the area between -1.96 and +1.96 equals 95%. A 90% confidence interval captures a smaller area and has a z-score of 1.645. A 99% confidence interval captures a larger area and has a z-score of 2.576. Thus, the larger the level of confidence, the larger the margin of error.

The remaining element in the formula is the standard deviation. While you have the freedom to choose your level of confidence and your sample size, you have no freedom to choose a standard deviation, as it is computed from the data, and it is what it is. Yet, it has an impact. If the data is naturally spread out, the confidence interval will be spread out more, and vice versa. If I teach a daytime undergraduate class at a university and an evening MBA class at the same university, and I asked you to guess the average age of my classes within one year, you would need some data first. If I told you the ages of 5 randomly chosen students from my daytime class was 19, 20, 19, 21, 19, you would have no problem in guessing the mean age within one year since the spread of ages is apparently very slim (and understandably so). Now if I told you the ages of some of my MBA students was 28, 43, 55, 32, 39, you wouldn't know where to begin to guess within one year (you probably wouldn't even feel comfortable guessing within five years). You clearly would need a lot more data to accomplish this task.

Confidence intervals can also be computed for proportions. They are essentially the same in that there is still a confidence level (translated into a z-score) and sample size, but there are no standard deviations. Whenever you see an election poll that reports a margin of error, you are looking at a confidence interval. For example, if the Republican candidate is shown to have 52% of the vote with a 3-point margin of error, it means that we are X% confident the candidate truly has between 49% and 55% of the vote. The 52% is based on a sample, but the confidence interval implies that if the entire population were polled, the results would likely be in that range. Our certainty that the truth is in that range is the level of confidence. Notice how often the newspapers report the margin of error but not the level of confidence!!! Anyway, a range of 49 – 55% means that the candidate might win or might lose, and the race should be labeled as undecided. You shouldn't focus on the fact that it leans toward the positive; if it is not clearly a victory or a loss, it is straddling the line and a decision should not be made. Making a conclusion based on being close is like stating a woman is almost pregnant (there is no “almost” about it). If the confidence interval were 51 – 55%, then you could state that the candidate is winning, and you are X% confident of it.

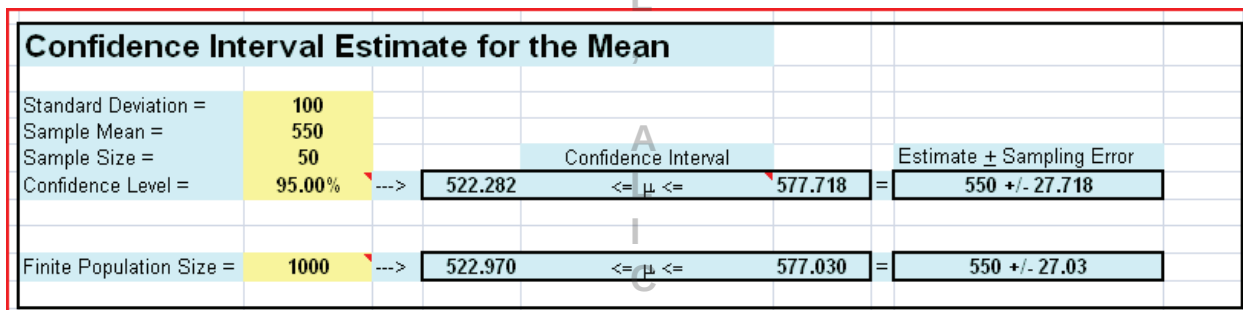
Let's look at how we can solve Confidence Intervals problems using Excel. Suppose a sample of 25 students at Whatsamatta U. was found to have a mean GMAT score of 550 with a standard deviation of 100. What is the 95% confidence interval of the mean GMAT score for all students at that university?

Figure 1: Screen display from **Confidence Interval.xls**



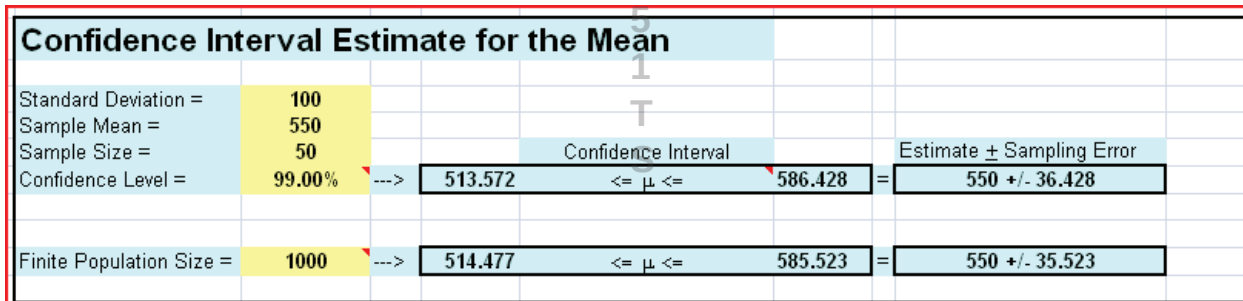
Plugging in the data appropriately in the **Confidence Interval.xls** file yields a 95% confidence interval of 508.722 – 591.278 or 550 $\pm$ 41.278. This means we are 95% confident that the population mean GMAT score at Whatsamatta U. is somewhere between 508.722 and 591.278. Now let's see what happens when we change some of the factors.

Figure 2: Screen display from **Confidence Interval.xls**



What if the mean was based on a larger sample size? Here you can see that a sample of 50 generated a tighter confidence interval, with a width of $\pm$ 27.718.

Figure 3: Screen display from **Confidence Interval.xls**



The larger level of confidence results in an increase in the sampling error.

So you can clearly see the effects of changes in the confidence level or sample size on the size of a confidence interval. If by chance there is a limited population size, such as only 1000 in the entire MBA program, plug that in to the FINITE POPULATION SIZE box and you will see an even tighter interval, depending on how small the population. Notice how a population of 1000 didn't really cut the interval size down much, and populations of 10,000 and up will likely show no difference, but very small populations will cut the interval significantly.

Figure 4: Screen display from **Confidence_Interval.xls**

Confidence Interval Estimate for the Mean					
Standard Deviation =	100				
Sample Mean =	550				
Sample Size =	50				
Confidence Level =	99.00%	--->	513.572	<= μ <=	586.428
					Estimate $\pm$ Sampling Error
					550 $\pm$ 36.428
Finite Population Size =	51	--->	544.848	<= μ <=	555.152
					550 $\pm$ 5.152

In Figure 4 above, a population size of 51 was entered, and the sampling error was cut to ± 5.152 ; note that at this point you have a sample which consists of 50 people out of the 51, so only one is missing, and cannot impact the mean by much.

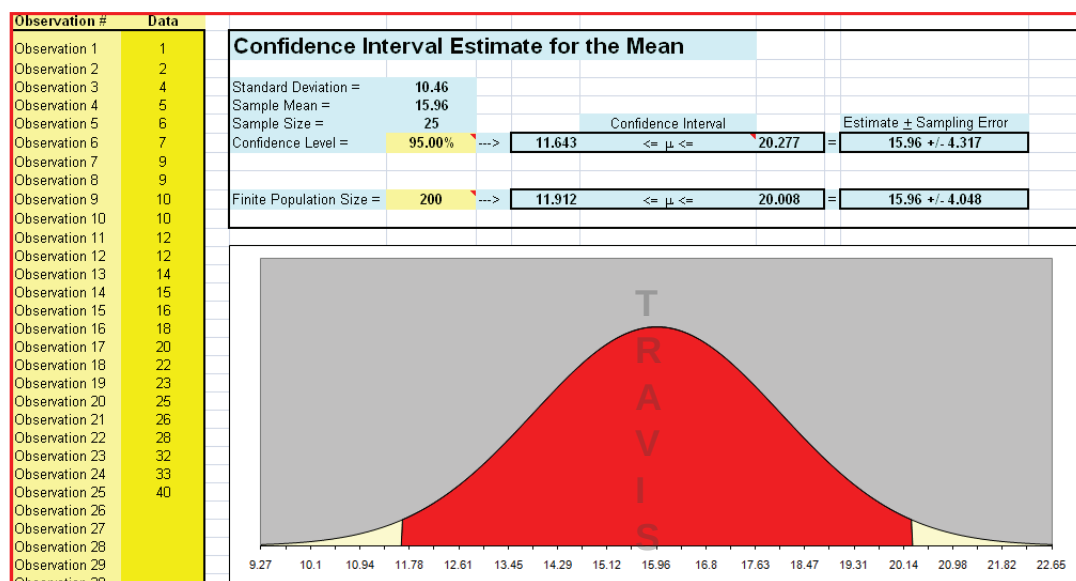
Figure 5: Screen display from **Confidence_Interval.xls**

Confidence Interval Estimate for the Mean					
Standard Deviation =	100				
Sample Mean =	550				
Sample Size =	50				
Confidence Level =	99.00%	--->	513.572	<= μ <=	586.428
					Estimate $\pm$ Sampling Error
					550 $\pm$ 36.428
Finite Population Size =	50	--->	550.000	<= μ <=	550.000
					550 $\pm$ 0

Note what happens when we enter a population size of 50, the same as the sample size. The sampling error is now zero, which makes sense since your sample equals the population, so there can be no sampling error, as you know the truth about the population.

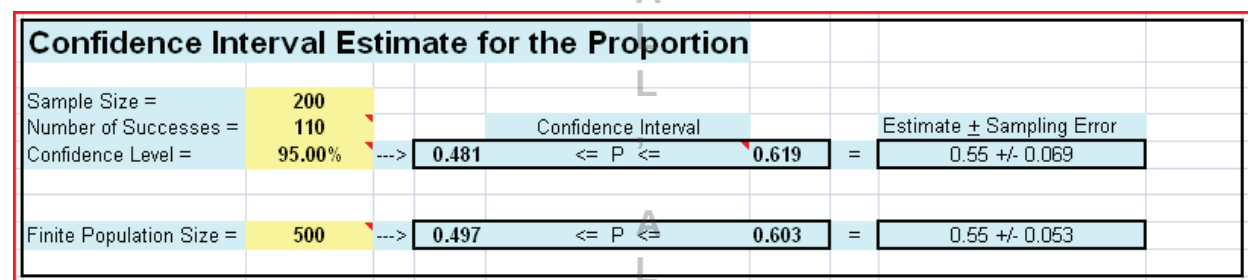
If you should have raw data, you can use the RAW DATA tab in the file which will compute the sample mean, standard deviation and sample size for you (as seen in Figure 6). You still choose the confidence level, and the rest is the same.

Figure 6: Screen display from **Confidence Interval.xls**



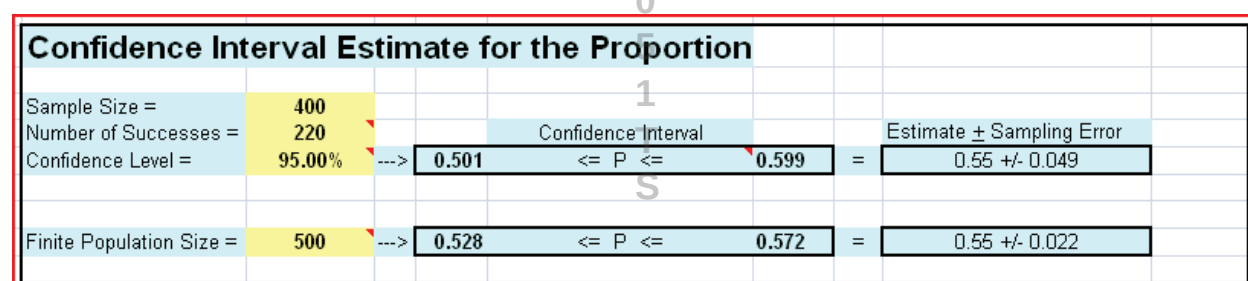
Now let's look at a proportion problem. If 200 people are polled and 110 reveal they voted for candidate X, that is 55%.

Figure 7: Screen display from **Confidence Interval.xls**



The 95% confidence interval shows a margin of error of .069, meaning that we are 95% confident that the true proportion is between 48.1% and 61.9%. We would not be able to conclusively declare the candidate a winner since he could be below or above 50%. The election would be undecided at this point.

Figure 8: Screen display from **Confidence Interval.xls**



Now if our sample were twice as large and we had 220 out of 400 declaring their vote for Candidate X, he still has 55% of the sample vote, but the margin of error is now 4.9%. Thus we are 95% confident his actual vote count will be somewhere between 50.1% and 59.9%, which means he will likely win since the entire interval is above 50%.

Sample Size

Probably the most common question often asked of a statistician is “how large of a sample do I need to take...” To understand the answer to this, it is important to look back a bit. When data is normally distributed, you can compute probabilities if only you knew the mean and standard deviation, and life is good. But what if you don’t know the population mean? In that case, the best estimate of a population mean is a sample mean. So take a sample and compute the mean and you have an estimate. But how good is that estimate? That’s what margins of error are for, based on your level of confidence. But what if you compute your confidence interval and it is just too large (after all, what good is it to estimate that the presidential candidate has 52% of the vote with a margin of error of 20%)? If the confidence interval is too large, you can always take a larger sample to tighten the confidence interval, but how much is enough? Wouldn’t it be nice to know exactly how large a sample you need to get the confidence interval you desire? Thanks to the sample size formula, that is possible.

To compute the minimum sample size, you need to answer two questions: how much error are you willing to tolerate AND how confident do you want to be about the truth being within that margin of error. The more confident you want to be, the larger the sample you’ll need. The greater the margin of error you can tolerate, the smaller the sample you will need. To illustrate, if I gave you a dart and asked you to hit the bullseye on the first throw, you probably wouldn’t bet on being successful. If asked how many darts you would need to throw before you were 90% confident you would hit the bullseye, you would request several. If asked how many it would take for you to be 99% confident of hitting the bullseye, you would request even more. The more confident you need to be of hitting the target, the larger the sample you would need. If asked about how close you would come to the bullseye with just one dart, you might ask for a lot of room for error. If given two darts you would probably feel better at getting closer to the bullseye. The more darts you throw, the closer you feel you will get to the bullseye. And so the closer you need to get to the bullseye (i.e., the less error you can tolerate), the larger the sample size you need.

Normally when asked about the tolerance for error and the confidence level, those not in the know is to say they want 0% error and 100% confidence, but that would require sampling the entire population. Anything less than the population involves incurring a degree of error, but the question is “how much is acceptable?” Once you decide on a margin of error and confidence level, the sample size is computed, but what if you need to sample 2,000 people? If you do a phone survey, you may have to call 20,000 people just to get 2,000 responses, but if you cut the sample size you also cut the confidence and increase the margin of error. It is truly a balancing game, but you are in control.

With a knowledge of confidence intervals and sample sizes, you now have an insider’s knowledge

of interpreting survey results, and what a margin of error truly means, and why we prefer the error to be small, and how larger sample sizes translate to higher levels of confidence with smaller margins of error. This is often a misused concept, especially by the media, but a little knowledge of it gives you an advantage over the majority of Americans.

Figure 9: Screen display from [Sample_Sizes.xls](#)

Sample Size Determination for Mean			
Population Standard Deviation =	100	T R A V I S I T A N T S	
Sampling Error =	25		
Confidence Level =	95.00%		Sample Size Needed = 62
Finite Population Size =	500		Sample Size Needed = 55

Now computing a minimum sample size is merely a shortcut to getting the confidence interval you desire. If you wished to know the mean GMAT score at a university within 25 points and with 95% confidence, according to this computation you would need 62. That means a sample of 62 GMAT scores would result in a 95% confidence interval with a 25-point sampling error.

Figure 10: Screen display from [Sample_Sizes.xls](#)

Sample Size Determination for Mean			
Population Standard Deviation =	100	L , A L I C E	
Sampling Error =	50		
Confidence Level =	95.00%		Sample Size Needed = 16
Finite Population Size =	500		Sample Size Needed = 15

If you were willing to tolerate more error, you wouldn't need as large a sample. Here the tolerable sampling error was doubled and the minimum sample size was reduced 75%.

Figure 11: Screen display from [Sample_Sizes.xls](#)

Sample Size Determination for Mean			
Population Standard Deviation =	100	S T U D E N T S	
Sampling Error =	40		
Confidence Level =	99.00%		Sample Size Needed = 42
Finite Population Size =	500		Sample Size Needed = 39

If you needed to be more confident in your findings, notice how the minimum sample size increases.

Figure 12: Screen display from **Confidence Interval.xls**

Confidence Interval Estimate for the Mean					
Standard Deviation =	100				
Sample Mean =	550				
Sample Size =	42				
Confidence Level =	99.00%	--->	510.254	<= μ <=	589.746 = Estimate $\pm$ Sampling Error
					550 $\pm$ 39.746
Finite Population Size =	200	--->	514.584	<= μ <=	585.416 = 550 $\pm$ 35.416

And so if you sampled 42 GMAT scores, and then created a 99% confidence interval of the mean, you would find it to have a sampling error of ± 40 points, which is demonstrated above.

Figure 13: Screen display from **Sample Sizes.xls**

Sample Size Determination for Proportion			
Estimate of True Proportion* =	0.50		
Sampling Error =	0.05		
Confidence Level =	95.00%		
		Sample Size Needed =	385
Finite Population Size =	500		
		Sample Size Needed =	218
*Use .50 for the "estimate of true proportion" when unknown.			

While procedurally the same as for means, sample size computations for proportions are more commonly performed, especially since polling is so popular. Suppose you wanted to conduct a poll and be 95% confident of capturing the truth within 5%, accordingly you would need to survey 385 people. And as with the means, an increase in tolerable error will decrease the sample size needed, and vice versa. And an increase in confidence levels will increase the sample size needed, and vice versa. Notice the ESTIMATE OF TRUE PROPORTION value of .50; if you have historical data to support what the true proportion probably is, enter it here. For example, if you are sampling for defective widgets, and historically you expect to find 12% defectives, then use .12 for this value; values here range from 0.00 to 1.00, and the further the value is from .50, the smaller the sample size is. In the absence of any information, use .50 as the default, as it will result in the largest possible sample.

As for the sample sizes for means, the results have the same implications. Here, if you take a sample of 385 people, your 95% confidence interval of the proportion will have a sampling error of $\pm 5\%$.

So now you have an inside knowledge into the methods used for election polls and what triggers the decision to declare a state as red, blue or undecided.

CHAPTER THREE KNOWLEDGE ASSESSMENT

Confidence Intervals & Sample Sizes

Discussion Questions

The results of voter surveys and polls guided the political campaigns of many presidents and their opponents and advised news groups on which political strategies are working. While polls have an important role in the political process, they are anything but perfect tools for measuring and forecasting voter behavior. Public opinion often changes daily or weekly (or so the polls report this to be the case). And even worse, competing polls report contradictory results over the same period of time. How are you to know which polls to trust, if any?

Look at <http://www.ncpp.org/?q=node/4> (20 Questions a Journalist Should Ask About Poll Results) and <http://www.pollingreport.com> (Polling Report). Answer the following two questions.

DISCUSSION QUESTION 1

SIZE VS. VALIDITY: In evaluating a poll, a larger sample size would be considered more favorably, but what consideration would you consider even important than the sample size before trusting the validity of a poll?

DISCUSSION QUESTION 2

POLLING RESULTS: Choose any ONE of the polls from Polling Report. Compare the many polls used to address a single question and explain possible differences in results.

Please don't turn this into a heated debate -- just have fun with it.

Practice Problems: Real Estate

Solutions are provided to practice problems so you can check your work.

Use the Real_Estate.xls file which consists of 100 homes purchased in 2007 and appraised in 2008. It includes variables regarding the number of bedrooms, number of bathrooms, whether the house has a pool or garage, the age, size and price of the home, what the house is constructed from, how far it is to the city center, and the appraisals from two agents.

PRACTICE PROBLEM 1: You wish to know the average sale price of a home in 2007. Compute the 95% confidence interval of the mean using the sample of 100 homes. Describe your findings.

PRACTICE PROBLEM 2: You wish to know the proportion of homes with a swimming pool in 2007. Compute the 95% confidence interval of the proportion using the sample of 100 homes. Describe your findings.

CHAPTER THREE KNOWLEDGE ASSESSMENT

Confidence Intervals & Sample Sizes

PRACTICE PROBLEM 3: You wish to learn the average appraised price of a home in 2008 within \$5000, and with 95% confidence, assuming a standard deviation of \$43,250. How large of a sample should you get? *(note: the standard deviation shown here was computed from the 200 appraisals in the sample; unless historical data is provided, the best option is to take an initial sample of at least 30 to get an estimated standard deviation).*

PRACTICE PROBLEM 4: You wish to learn the proportion of homes with a garage within 3%, and with 90% confidence. How large of a sample should you get? *(note: the estimate of the true proportion is defaulted to 0.50 unless you have prior knowledge or an initial sample. Since there is a sample of 100 in the data file and 66 of the 100 have a garage, you can use 0.66 as the estimate. standard deviation shown here was computed from the 200 appraisals in the sample; unless historical data is provided, the best option is to take an initial sample of at least 30 to get an estimated standard deviation).*

Assigned Problems: Student Data

Use the Student_Data.xls file which consists of 200 MBA at Whatsamattu U. It includes variables regarding their age, gender, major, GPA, Bachelors GPA, course load, English speaking status, family, weekly hours spent studying.

ASSIGNED PROBLEM 1: You wish to know the average GPA of MBA students at Whatsamattu U. Compute the 95% confidence interval of the mean using the sample of 200 students. Describe your findings.

ASSIGNED PROBLEM 2: You wish to know the proportion of MBA students that are majoring in Finance. Compute the 95% confidence interval of the proportion using the sample of 200 students. Describe your findings.

ASSIGNED PROBLEM 3: You wish to learn the average age of an MBA student within 2 years and with 99% confidence. How large of a sample should you get?

ASSIGNED PROBLEM 4: You wish to learn the proportion of MBA students that are female within 3%, and with 98% confidence. How large of a sample should you get?