

## **Survey Question Analysis (Draft 2008-12-19)**

The purpose of this tutorial is to *analyze* two types of questions commonly found on surveys: proportion (yes/no) questions and Likert scale (preferences) questions. (This tutorial doesn't tell you how to 'design' your questionnaire, but I do list a few types of questions you might want to ask, see "Survey Pointers" on page 12). I describe very simple calculations for sample sizes, error intervals, and confidence intervals associated with these survey questions. I have also included calculations for *small* populations where your sample size is an appreciable fraction of the population.

After working through this tutorial you will be able to determine what errors you can anticipate by sampling, say 50 people, and in turn, how to determine how many people you need in order to achieve a specified error size of say, 10%. (for example, if it's a yes/no question, then in the first case you are going to be off by a little more than 14%. To get down to a 10% error you will need about 100 responses - always with the caution that your sample of respondents represent the population).

.For example, for *proportion* responses: if 40% of your sample respondents say "yes" and 60% say "no" to a question on a sample survey, what confidence can you place on this 40% number? You will see that this depends on the number of people surveyed, and, you can actually state how confident you are that this 40% ( $\pm$  an error), represents the true proportion who would say yes, if you could ask everyone. For example, if you sampled 50 representative people then you could claim that the true proportion was  $40\% \pm 14.1\%$ , at a 95% confidence level. That is, the true proportion of people saying yes would be in the range from about 26% to 54%, 95 times out of 100 if the identical experiment were performed 100 times.

### **Cut to the Chase**

I have given a very simple way to calculate both the error and confidence interval for proportion questions, for a given sample size, and the necessary sample size to achieve a given error and confidence interval.

*When you know Your Sample, Size, 'n', What is the Error for Proportional Sampling?*

**Error =  $1/\text{Sqrt}[n]$ ,**

that is, if you plug in your number of samples, 'n', you can calculate the error made, at a 95% confidence level. So, for  $n=50$  respondents, the error percentage is:

$1/\text{Sqrt}[50] = 0.1412 = 14.1\%$  with 95% confidence

*When You Know What Error 'E' You Will Tolerate, What Must 'n' Be?*

**$n = 1/E^2$**

that is, if you can only tolerate a certain error, then plug in your tolerated error value 'E' (in decimal form), then you can calculate what your sample number of respondents must be. So, if I will tolerate an error of 10%, (0.10), the minimum sample size would be,  $n = 1/(0.1)^2 = 100$ .

Similarly, I will look at a survey that includes questions of the *Likert* type scale and then

show how to calculate the confidence interval that may be placed around that average outcome, for a given question.

These next sections elaborate on the Cut to the Chase section above as well as discussing analyses of Likert scales.

## Two Survey Types and Their Analyses

What follows is an expansion of the Cut to the Chase section above.

### “Proportion” Surveys

In these surveys I am going to get *proportion/percentage response* from a *sample* yes/no questionnaire, and my interest is to estimate the *true population percentage*, together with an associated “error”. Taking this a little further, I would like to be able to calculate *how many* respondents I need to sample if I want to be sure of getting some specified maximum error. A Maximum Error means an interval around the proportion that I actually observed, that will contain the true proportion, a specified percentage of the time. In other words, I take a sample and get an actual proportion (which is almost certainly not the true proportion). Now I want to know where the true (population) proportion lies? Knowing the number of responses, I can actually calculate an interval around the sample proportion and be pretty sure (*at a specified confidence level*) that the true population proportion lies somewhere in that interval.

### Background for Calculating Proportion Confidence Intervals

Suppose I am interested in the number of people who say that they voted *yes* on proposition X, in a recent election. Suppose I cook up such a yes/no questionnaire, go to the election exit area, and ask 50 people whether they voted yes or no on proposition X. Why 50 people? For this example I am assuming that I happened to have the time and energy to ask only 50 people. Further, I don’t really know, but I assume that these 50 people are representative of the population I want to find out about. In other words, I didn’t initially consider how that number would quantitatively determine my result accuracy. (we will consider that a little later).

Suppose that 20 out of the 50 said yes and 30 said no. The *sample* proportion who said “yes”, which is what I actually observed, is  $20/50 = 0.4$  or 40% The percentage who said “no” then was 60% or a proportion of 0.6. The figure below shows the actual outcome noted as 0.40. What is the true actual proportion? I have shown this as  $p^{\text{true}}$  and have also indicated that I don’t really know where it is. I couldn’t know  $p^{\text{true}}$  unless I asked every member of the whole population. Normally, you won’t have access to the total population, and so, you are going to have to take a sample.

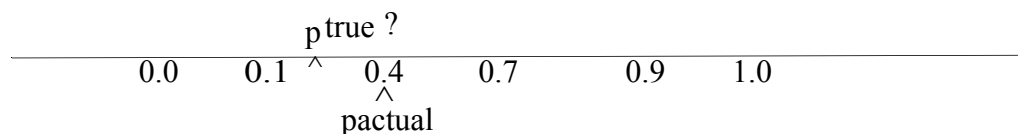


FIGURE 1. Actual sample proportion and unknown true proportion

### **\*Quick Start Calculation for Error Made for a Given Sample Size**

Caveats for calculations: These quick start calculations make the following simplifications: Replace the exact z value for a 95% confidence interval, which is “1.96”, with a value of “2”. Use the most conservative estimate of proportional standard deviations, by using “1/2” for p and (1-p). The result is a 95.45% confidence interval and a slightly larger sample size.

If you are constrained to take a representative sample of, say “n” people, then the error you will make, at the 95% confidence level is, at most,  $\text{Sqrt}[1/n]$ . This is a very good approximation to “2 sigmas”, that is, two standard deviations. Because of the math, factors cancelled out to yield this simple result:

#### **Error = $1/\text{Sqrt}[n]$**

For example, if you could only sample 50 people, then the maximum error you would make, at the 95% confidence level would be:

$$1/\text{Sqrt}[n] = 1/\text{Sqrt}[50] = 0.1414 \text{ or a little more than } 14\%$$

#### **The confidence interval is your sample outcome proportion $\pm$ Error**

For example, if your sample outcome proportion was 40%, then the 95% confidence interval is  $0.4 \pm 0.14$  which is from 26% to 54%. This means that the *true* proportion will lie in this interval 95 times out of 100, if you randomly sampled that often, under otherwise identical conditions!

### **\*Quick Start Calculation for Sample Size Needed for a Desired Error**

If you want to keep the error below a given percentage, then you adjust the sample size. The formula below will give you a 95% confidence level for the specified error.

$$n = 1/\text{Error}^2$$

For example, if you can only tolerate at most a 5% error in your survey, then your sample size should be at least:

$$\begin{aligned} n &= 1/(0.05)^2 \\ &= 400 \end{aligned}$$

This means that if your sample of 400 people answered a given question ‘yes’ 40% of the time, then you could report that the true population proportion was  $40\% \pm 5\%$ , with 95% confidence.

\*\*If you would like to look a little more closely at how these values are calculated and some background on their use, check out the section later in this tutorial, “For Those Interested in the Basis for the Proportion Quick Calculations” on page 9.

### **When the Population is Small, You Can Do More With Less**

Whenever your total population is on the order of 20 times or more of your sample size then it makes sense to use the techniques just presented above. If though, your population is *less* than 20 times your sample size, then you can make stronger claims, since you are actually sampling a smaller population relative to your sample size.

O.k., suppose you are doing sampling from a small population. This means that you are only going to generalize from your same results to this particular population. Although you can't get to everyone, you can get an appreciable portion of the population. If that is the case

then using small sample statistics allows you to claim a lower error rate as well as take smaller samples for the same error rate.

For example, we found earlier that 50 samples led to an 14.14% error rate. (this is the very large population context)

However, if the total population were only 200 people, then the error rate could be reported as: 12%. This is not too big an improvement, but it gets lots better very quickly. If you could get 100 of the 200 total, then you could report 7%! From the equations below you will be able to calculate these desired error rates and sample sizes.

**\*Quick Calculation for the Sampling Error from a Finite Population**

Given the discussion above, and the constraint that your population is less than 20 times your sample size, do the following:

let “n” be your sample size and “N” is the size of your known population. Note that this calculation is the same error calculation as above with the additional finite population correction factor,  $\text{Sqrt}[1-n/N]$ .

**Error =  $1/\text{Sqrt}[n] * \text{Sqrt}[1 - n/N]$**

For example, if you again sampled 50 people as in the example above but, now you know that the size of the population is a finite number like 200, your error rate will be rather than the 14.14% calculated earlier. In other words, it takes that 14% and scales it down by the *finite population correction factor*  $\text{Sqrt}[1 - n/N]$ .

error =  $\text{Sqrt}[1/50] * \text{Sqrt}[1 - 50/200] = 12\%$

If you could sample 100 out of the 200 population, then the error would go down to:

error =  $\text{Sqrt}[1/100] * \text{Sqrt}[1 - 100/200] = 7\%$

**\*Quick Calculation for the Sample Size for Finite Population, for a Specified Error**

If you want to keep the error below a given percentage, then you adjust the sample size. The formula below will give you a 95% confidence level for the specified error. Compare this result with the earlier result for a very large population relative to the sample size.

**$n = N/(1 + \text{Error}^2 * N)$**

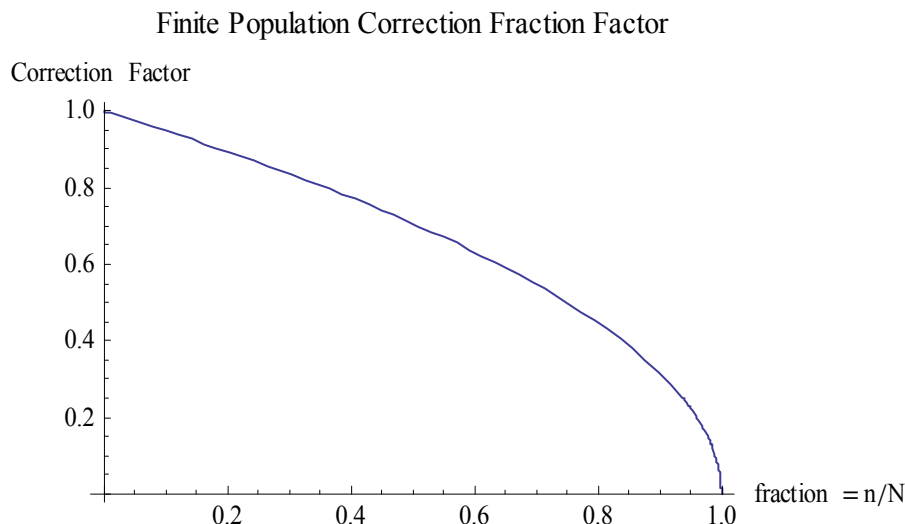
For example, if you would like to hold your errors to 5% or less and the total population was 200, then you would need at least:

$n = 200/(1 + 0.05^2 * 200) = 134$

(Compare this with the general case where you would need:  $n = 1/.05^2 = 400$ )

*Just for Fun, Here is a Plot of the Finite Population Correction Factor*

This plot shows how the correction factor, which scales the error downwards, depends on the ratio of sample size to total population. For example, when the sample size is approximately half the size of the population, then the scale factor is about 0.7. This means that the error is reduced by some 30% from its ‘large population’ value.



**FIGURE 2. Correction Factor Versus the Ratio of Sample Size to Total Population**

### Likert Scales (Preference Scales)

This is the most popular interval scale, and involves assigning numerics, usually 1 to 5, or 1 to 7, or perhaps 1 to 9, to each question, corresponding to differing degrees of agreement with the question. Actually, you don't have to show your respondents the numbers, just the range of questions. Later, when you are analyzing the results, you map the answers to numbers and go from there. You have probably encountered this type of survey if you have done an end-of-course/training/hospital stay/job satisfaction survey.

Below is an example from a Likert type survey assessing the statistical sophistication within an organization. I have also included the scale, 1-5, here but that's not necessary.

Question 1 (Circle the number that is closest, in your opinion)

“Statistical Methods are Currently Widely Used in my Organization”

Strongly		Neither Agree		Strongly
Agree	Agree	nor Disagree	Disagree	Disagree
5	4	3	2	1

**FIGURE 3. A Likert type question**

#### Analyzing the Likert Question

For this question, I am going to calculate an overall mean of responses and try to infer what the total population would have answered, on average. Keep in mind though, that this approach smears out highs and lows. This may or may not be what you want. Another perspective is to look individually at the high ratings and the low ratings. Doing multiple analyses is usually the best approach to take and is in keeping with the ‘explore first and then decide what to try to confirm’ attitude.

Assume you had 10 respondents to the above question, with the distribution of responses

## Survey Question Analysis (Draft 2008-12-1) Likert Scales (Preference Scales)

as follows:

Agreement	Number of Respondents answering with that response.
Strongly Agree: #5	2
Agree: #4	3
Neither Agree nor Disagree: #3	3
Disagree: #2	1
Strongly Disagree: #1	1

One person answered with a “1”, one person answered with a “2”, three people answered “3”, and so on. Let me write this set of responses as a vector, the *observation vector or outcome vector*.

This shows there was a single ‘1’, a single ‘2’, three ‘3’s, three ‘4’s, and two ‘5’s.

observed  $x = \{1, 2, 3, 3, 3, 4, 4, 4, 5, 5\}$

$n = 10$ , sample size, that is, there were 10 respondents to this question

sample mean “ $\bar{x}$ ” = 3.4

sample variance =  $\sum (x_i - \bar{x})^2 / 9 = 1.6$  (see the rationale for using ‘9’ below), where “ $i$ ” takes on the values of, 1 through 10.

sample standard deviation ( $s$ ) =  $\text{Sqrt}[\text{sample variance}] = 1.26$

Note that I divided by “9” above rather than “10”. This is due to a statistical theorem that says that using this denominator, rather than “10”, will give a closer estimate to the true variance,  $\sigma^2$ , in the long run.

### What is the True Population Mean?

We have taken one sample and gotten the results above, where  $\bar{x}$ , the sample mean was 3.4. If we took another sample of 10 people, we would get a different set of outcomes, and so a different  $\bar{x}$ . And if we kept doing this, over and over, we would build up a whole set of “ $\bar{x}$ s” and their associated standard deviations. It turns out that the distribution of *the set of  $\bar{x}$ s* is more closely packed than any individual sample set of values. That is, if we collect  $\bar{x}$ ’s, and look at how those group together, they will group more closely than the values in an individual sample. If we plot the  $\bar{x}$  values, we will see that a graph of them starts to look like a bell shaped (normal) curve. In the long run, the  $\bar{x}$ s will tend to the true population mean and the variance of the  $\bar{x}$ s will tend to the true population variance, scaled by ‘ $n$ ’.

### Standard Error of the Mean

The *standard error of the mean* measures the standard deviation of the  $\bar{x}$ ’s. That is, the standard error of the mean describes the spread of the (potentially obtained) *sample means*.

If we knew the actual population standard deviation, denoted by sigma  $\sigma$ , the standard error of the mean could be written as:

Standard Error of Mean =  $\sigma / \text{Sqrt}[n]$

If we don’t know the true population standard deviation (and we usually don’t) we use the sample standard deviation,  $s$ , that we calculate from the sample. Doing this results in:

Standard Error of Mean = sample standard deviation (s) / Sqrt[n]

In the example above, we have, for a particular outcome:

n = 10

xbar = 3.4

sample Variance = 1.6

s = sample standard deviation = 1.26

so, the standard error of the mean will be

Standard Error of Mean = s / Sqrt[10] = 0.39

### **Estimating the Population Mean (Constructing a Confidence Interval)**

The value of using the standard error of the mean comes when we want to try to pin down the location of the true mean. This will be analogous to calculating the proportion confidence interval as we did earlier.

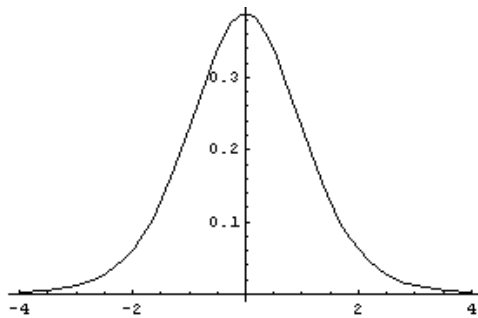
Technically, when the sample size is less than 30, and you don't know the population standard deviation,  $\sigma$ , there is a more specialized reference distribution that is used to perform significance tests and calculate confidence intervals. That is called *Student's T Distribution*, or the T Distribution for short. For samples of 30 or more, it matches more and more closely the Normal Distribution, but for smaller samples, it is a little more spread out than the Normal (fatter tails) and is keyed to the actual sample size. Note that although the sample size is "10", the distribution is shown with a "9". If you recall your stats, this is the *degrees of freedom* of that sample and is "n-1 = 10-1 = 9".

Degrees of freedom (df), in this case, refers to the dimension of the *error space*. For a sample size of 10, the observation vector, 'x', sits in a 10-dimensional space. In this particular case, the error space is of dimension 9, while the dimension of the model space is "1", for a total of 10. If you are looking for the mean of a single population, then the error space is going to be the sample size "n" minus 1, that is, n-1=9. The model space is the space that contains the true mean, in the form of a one dimensional vector.

```

sdist = StudentTDistribution[9]
studentPDF = PDF[sdist, x]
Out[24]=  $\frac{2519424}{35 \pi (9 + x^2)^5}$ 
Plot[studentPDF, {x, -4, 4}]
In[31]:= t95 = Quantile[sdist, 0.975]
Out[31]= 2.26216
In[34]:= -t95 = Quantile[sdist, 0.025]
Out[34]= -2.26216
In[35]:= Integrate[sPDF, {x, -t95, t95}]
Out[35]= 0.95 + 0. i

```



The program Mathematica has built in stat functions. Some parameters of the T distribution are shown on the left in this diagram. The number "9" means that this member of the T distribution is tuned to sample sizes of 10 (where the "9" is what is called the degrees of freedom of a sample of size 10). The PDF is the distribution of values for the reference curve. Notice that it is symmetric since the  $(9 + x^2)$  factor is always positive and symmetric for both + and - x values. Further, when x gets large the PDF gets small, and so the tails of the curve go to the abscissa, just like the Normal curve.

The critical values such that 95% of the curve lie between them are at  $t_{95} = 2.262$  and  $-t_{95} = -2.262$ . I use these to calculate a confidence interval that captures the true mean 95% of the time. The "Quantile[sdist,0.975]" is a built in function that finds the t value such that 0.975 of the curve lies to its left, and this is the number 2.262. To confirm this, I integrated between the critical values to verify that 95% of the curve lay between them. (You can ignore the zero contribution of the complex component of this number)

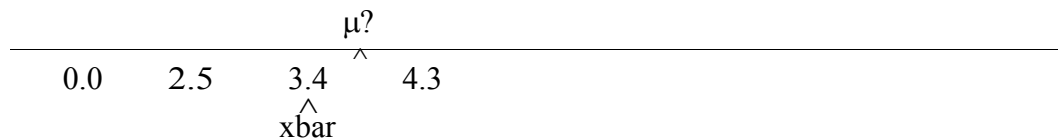
FIGURE 4. T- Distribution for Small Sample Calculations, Degrees of Freedom = 9

### Calculating the Confidence Interval around the Sample Mean (for small sample sizes <30)

Now we have the machinery to calculate a confidence interval around the sample mean, 3.4, that we predict will contain the true mean,  $\mu$ , 95% of the time.

Notice that because the sample is small, and we don't know the true variability, our interval of confidence will be larger than in the large sample case. This shows up in having to use the value  $t_{95} = 2.226$  (for 9 degrees of freedom) rather than  $z_{95} = 1.96$  (this 'z' value is independent of the sample size, so long as it is more than, say, about 30). The equation below shows the true mean to be pinned down between 2.5 and 4.3. This says that the true average response to the questionnaire, given I asked only 10 people, is somewhere between 2.5, and 4.3

$$\begin{aligned}
 \bar{x} - t_{95} * s/\text{Sqrt}[10] < \mu < \bar{x} + t_{95} * s/\text{Sqrt}[10] \\
 3.4 - 2.262 * 0.4 < \mu < 3.4 + 2.262 * 0.4 \\
 2.5 < \mu < 4.3
 \end{aligned}$$



### If That's Not Good Enough - Calculating the Sample Size to Achieve a Given Error

As it stands now, we know  $\mu$  to within a distance of  $t_{95} * 2.262 = 0.90$

If that is too much leeway, we can specify the required error interval and then solve for "n" just as we did for the proportion calculation. Suppose I want the true mean to be within 0.2 of the sample outcome mean,  $\bar{y}$ .

$$|\bar{x} - \mu| = 0.2$$

$$0.2 = t_{95} s / \sqrt{n}$$

$$n = (t_{95} * s / 0.2)^2$$

$$n = 204.$$

We can do a little better by recognizing that if the sample size is  $>30$  we can use the Normal curve critical values,  $z_{95}$  and  $-z_{95}$ . Plugging those in yields:

$$0.2 = z_{95} * s / \sqrt{n}$$

$$n = 154$$

### Larger Sample Sizes Allow Use of the Normal Curve

If we sample more people, say, more than 30, then we could use the Normal curve values such as  $\pm z_{95}$  and so get a tighter confidence interval.

### For Those Interested in the Basis for the Proportion Quick Calculations

In the discussion below, instead of the "textbook" values for the critical "z" value for 95% confidence, which is 1.96, I use the value of 2.00. I am referring here to the common use of  $z = \pm 1.96$  to include exactly 95% of the normal curve as opposed to using the value " $\pm 2.00$ " which is easier to work with and includes 95.45% of the normal curve.

Let me use the following notation:

$p^{\text{true}}$  = proportion to be estimated (this is the true proportion of the whole population of those saying yes, if you could question every person exiting the voting area.)

$p_{\text{observed}}$  = proportion you actually get as a result of your sample. (Note: this is called a "point estimate" of the true proportion). I am assuming that I found this to be 0.4

So the observed proportion who voted yes is:  $p_{\text{observed}} = 0.4$

$n$  = sample size (total number of people I asked) = 50

$\text{var}$  = Variance of a proportion (estimated) =  $p_{\text{observed}} * (1 - p_{\text{observed}}) / n$

(This variance formula can be derived, but for now, just accept it! o.k?)

NOTE: to get the most conservative estimates, replace  $p_{\text{observed}}$  and  $(1 - p_{\text{observed}})$  with 0.5. That is, using a 'p' value of 0.5 will increase the sample size slightly but also increase confidence. I will use this more conservative estimate in subsequent calculations.

## Survey Question Analysis (Draft 2008-12-1) Likert Scales (Preference Scales)

std = Standard deviation of a proportion = Sqrt [Variance of proportion] =

$$\text{Sqrt } [0.5*(1-.5) / 50] = 0.0707$$

$z_{95+} = 2.00$  (this critical value is chosen by the experimenter to achieve a certain “confidence interval, in this case a 95+% confidence interval- see the discussion below). This also means that I am going to use the Normal approximation for these proportion estimates.

From this data I can calculate a *confidence interval* around the observed outcome proportion of 0.4, and assert with 95% confidence that the true proportion,  $p^{\text{true}}$  lies in the interval shown below: This means that I am going 2 standard deviations on either side of the observed outcome and claim that interval will include the true proportion 95% of the time.

$$0.4 - 2.00 * 0.071 < p < 0.4 + 2.00 * 0.071$$

$$0.26 < p < .54$$

The number  $2.00 * 0.071 = .142$  is the error, or about 14%

This tells me that I have about a 0.14 error, that is, a 14% error in estimating the true proportion of those who said yes.

When I report my results, I would say that 40% *say* they voted yes on proposition X, and that value has an accompanying error of  $\pm 13.7\%$ . I would further say that I can be confident 95% of the time that the true proportion of those who would say yes, would be found in the interval [0.26 \_\_\_\_ 0.54]

If I want to reduce the error, I will need to take more samples, that is, question more voters. How many to take is the topic of the section “Calculating the Sample Size for a Proportion Estimate” on page 5.

Based on this, my initial conclusion would be that people will vote yes 40% of the time with a 14% margin of error. My next step would be to take a larger sample and pin down the estimate of those who claim to have said “yes”.

### Notes on the Confidence Interval Calculations

To relate what I am doing to the standard textbooks, they will use the value of  $z_{95} = 1.96$ , rather than “2.00”, as the *critical* value of a Normal reference curve. The reason is that this  $\pm$  this value gives exactly 95% of the normal curve while  $\pm 2.00$  includes 95.45% of the standard normal curve.

For example, the reference curve we will be using, the Normal Distribution, is shown below in “The Standard Normal Curve and Its Critical Values” on page 12. Note that the total area under this standard curve is exactly 1.000.

From the curve shown, or from standard tables, the critical values that bracket symmetrically 95% of the curve are at  $\pm 1.96$ . That is, 97.5% of the curve lies to the left of  $z_{95} = 1.96$  while 0.025% of the curve lies to the left of  $-z_{95} = -1.96$ . The net result is that the area between  $-z_{95}$  and  $+z_{95}$  is 95%.

If you do your observations and come up with a sample proportion, you will want to report on an associated “confidence interval”. This interval of numbers is calculated so as to include the true proportion a certain percentage of the time, if you could repeat the exact same experimental protocol over and over. There are some standard confidence percentages, such as 90%, 95%, and 99%. For our work, we will use the 95% confidence level. This means that 95 times out of 100 repetitions of the exact same experiment, the true proportion will be found within the calculated “confidence interval”.

## *Survey Question Analysis (Draft 2008-12-19) Likert Scales (Preference Scales)*

NOTE: When you listen to percentage/proportion figures quoted in the news, the implicit assumption is that the error figure given is based on a 95% confidence interval. So, for example, if the news person says a survey reveals that 38% of voters will approve proposition X, with a 4% margin of error, the usual interpretation is that the true proportion that will vote for X lies in the interval [34% \_ 42%], 95% of the time.

The connection between this standard reference curve and the particular results we came up with is that the standard curve's abscissa is marked off in terms of standard deviations. For the standard reference curve, the value of one standard deviation is "1". So, what you are seeing with the number 2.00 on the standard curve is the location of "2.00" standard deviations away from the mean, which is zero in the standard case. To be able to use this standard curve in our particular case, we need to make a translation between our proportion (mean) = 0.4 and standard deviation = 0.07, and the standard mean = 0.0 and standard deviation = 1.00. The connection is that we map our mean to 0.4 and our standard deviation to 0.07. The result is that our values along the abscissa that include 95+% of the data have 0.4 as the mean and  $0.4 \pm 2.00 * 0.07$  as the critical values.

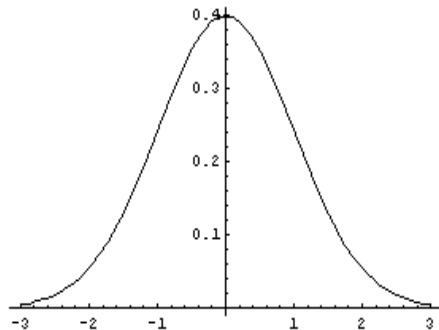
The graph below shows the standard normal curve with the values of 1.96 distinguished. Although I have used 2.00 in my calculations, the 1.96 is the one you will see in most textbooks.

```

ndist = NormalDistribution[0, 1] (* built in stat function *)
normalPDF = PDF[ndist, x] (* the density function *)
Out[34]= 
$$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

Plot[normalPDF, {x, -3, 3}] (* plot the density function *)
Quantile[ndist, 0.975] (* 97.5% of the area of curve
lies to the left of this value *)
Out[36]= 1.95996
Quantile[ndist, 0.025] (* 2.5% lies to the left of this value *)
Out[37]= -1.95996
NIntegrate[normalPDF, {x, -Infinity, 1.95996}]
(* check calculation, 97.5% is area to the left of 1.95996 *)
Out[38]= 0.975

```



The *Mathematica* package has many stat functions built in as well as their associated functions. For the Normal distribution, I have chosen the standard reference distribution with mean = 0 and standard deviation = 1. This distribution is highly useful as an approximation to the distribution of outcomes of many types of experiments/observations. That is, often the outcomes form a ‘bell shaped’ curve.

PDF[ndist,x] means the probability density function and shows the density of values as you go from left to right, being most dense around zero.

The total area under the Normal curve is “1” or 100%. If you check out the area under the curve between -1.96 and +1.96 you will find it to be 95%.

“Quantile” is a built in function that allows a quick answer to “what is the value such that ‘x’ percent of the curve lies to its left”. For example, the Quantile[ndist, 0.975] shows this value (called a critical value) to be 1.95996. This is the same result as integrating up to this value, starting at negative Infinity!

FIGURE 5. The Standard Normal Curve and Its Critical Values

## Survey Pointers

I would like to break the survey down into three sections, demographics, psychographics, and open ended/ interview questions. The idea of asking the demographic questions is often to find categories of respondents that can then be further analyzed or intercompared.

### Demographics

Age

Gender

Education level

Income

Place of origin (ethnicity)

Home location

Language preferred

### **Additional Demographic Questions Depending of Specific Research Question Focus**

Next are just a few questions taken from several surveys focused on different objectives and so are not generally applicable.

Home owner/renter?

Marital status:

- a) Single
- b) Married/Domestic Partner
- c) Widow
- d) Widower

Members in the family

Children in the family

How many credit cards? 1-3 4-6, more

Debts if any?

Computer literacy

- a) Online banking/shopping/surfing/ info gathering

Disaster preparedness – profiles

Job description: Executive/ clerk/ manager

Owner/ operator/partner

### **Psychographics**

These are the questions that bring out the opinions of your respondents. Since they are opinions, they are (or ought to be) expressed as degrees of agreement or disagreement. What is often lacking in surveys is an explicit context in which the question is asked. Since opinions vary depending on context, it would seem reasonable to make this explicit. Finally, there needs to be a way to combine different questions together in order to express not only preference orders, but intensities of preferences.

(Opinions, Preferences, Like and Dislikes, Ranks, Venting, Feelings)

AHP scale - Recommended - 9 point scale (1, 2, 3, 4, 5, 6, 7, 8, 9) or N/A

## **Summary**

Let me stop here after the analysis of a proportion problem and a Likert scale type problem. I have introduced some simplified calculations for sample size, error bounds and confidence intervals. For small samples the T-Distribution was introduced and used to calculate sample sizes and confidence intervals.

Although I didn't discuss it in this tutorial, you are encouraged to follow up these studies with another technology called the Analytic Hierarchy Process. I consider the AHP to be a natural evolution beyond the usual survey methods when complex decisions must be made and *justified*. I have written a separate manual on the use of the Analytic Hierarchy Process that the student may refer to if interested.

## **References**

Cooper, D., P. (2006) Schindler, *Business Research Methods* 9th ed. McGraw-Hill, New York, ISBN 0-07-297923-2

Saaty, Thomas, (1980) *The Analytic Hierarchy Process*, Addison Wesley, New York.