UNDERSTANDING CONFIDENCE INTERVALS

Preface: This document employs the construct of political polling to explain the concept of a single confidence interval and its interpretation.

OVERVIEW OF SURVEY SAMPLING AND POLLING

Broadly speaking, a survey is a way of learning about a group of people by having some of the people in the group answer questions.

Obtaining a sample (i.e., sampling) is a tool used to estimate the values of population characteristics when a census of an entire population is not practical. A sample survey is a survey that is carried out using a sample of people who are intended to represent a larger population.

In this light, polls are sample surveys designed to figure out the opinions and preferences of the entire population without having to ask every single person what they think. In other words, the goal is to poll a subset of the entire population (this subset is a sample) and come up with an answer that is representative of what the population as a whole believes.

For example, if we wanted to know more about the opinions of the one million adults who live in a particular city, we might select a sample of 900 individuals from those one million people. A sample survey is then conducted using those 900 people.

The underlying mathematics used to generate confidence intervals are based on the notion of random sampling. However, when dealing with polls we are faced with numerous biases associated with the people who were questioned, as well as the nature of their responses. In this vein, true randomness is problematic at best. Hence, the notion of a random sample is replaced by a seeking a representative sample. Accordingly, we endeavor to obtain a sample that represents the characteristics of the entire population with respect to the question(s) being asked.

Two main questions associated with poll results are: “What is a ‘reasonable’ estimate?” and “How confident are you in this estimate?” The answers to these questions are found by examining the construct of a confidence interval.

The results of a poll are reported in two parts… the point estimate (aka the estimate) and the margin of error. Additionally, both the point estimate and the margin of error will differ depending on the sample; in other words, each sample will yield results that differ slightly.

For polling results, the point estimate is a single value that estimates the true population percentage associated with the poll question. A sample point estimate almost always differs from the actual but unknown population value.

The margin of error is an interval (i.e., range of values) that accounts for the sampling error; this error is inherent in the results of a survey. The sampling error occurs because we are observing a sample instead of the whole population. Simply stated, the sampling error accounts for the difference between the sample statistics (i.e., the point estimate and margin of error generated from a sample) and the true population parameter.

The margin of error is typically expressed using the ± symbol (read as “plus or minus”). Additionally, the margin of error has an associated level of confidence; this is a measure of accuracy and represents the probability that the confidence interval has captured (i.e., contains) the true population value. Confidence level values vary and typically range from 50% to 99%, with the most common value being 95%.

A confidence interval can be more easily understood by working through a few examples.

SCENARIO: A reporter claims that, based on the results of a recent poll, a majority of voters favor Ballot Proposal C.

Note: Since this claim is based on a polling result, there is an implied confidence level. This confidence level is typically 95%, but is almost never reported.

Note: A majority is defined as anything greater than 50%; consequently, a value of 50% is not a majority. This is because, if there are only two outcomes, a value of 50% would result in a tie.

Example #1

A poll (sample) of 400 likely voters yields the following:

Ballot Proposal C is Favored by 54% of Voters (this is a point estimate).

The poll has a margin of error (MOE) of ± 5%

This means…

The "true" percentage of voters who favor ballot proposal C is between 49% and 59%.

Graphically we have:

Range of Percentages for Voters Favoring Ballot Proposal C

```
<---------------------------------------(-------|-------)-----------------------------------> 
                                        49      54      59
```

Since a portion of the confidence interval is below 50%, the reporter cannot correctly claim that a majority of voters favor proposal C.

In other words, we conclude that there is not statistical evidence, at the 95% confidence level, that a majority of voters favor proposal C.

Example #2

A poll (sample) of 1,000 likely voters yields the following:

Ballot Proposal C Is Favored By 54% Of Voters (this is a point estimate).

The poll has a margin of error (MOE) of ± 3%

This means…

The “true” percentage of voters who favor ballot proposal C is between 51% and 57%.

Graphically we have:

Range of Percentages for Voters Favoring Ballot Proposal C

<---(--------|-------)--------------------------------->

51 54 57

Since <u>no</u> portion of the confidence interval is at or below 50%, the reporter <u>can</u> correctly claim that a majority of voters favor proposal C.

In other words, we conclude that there is statistical evidence, at the 95% confidence level, that a majority of voters favor proposal C.

<u>Example #3</u>

A poll (sample) of 1,000 likely voters yields the following:

Ballot Proposal C Is Favored By 53% Of Voters (this is a point estimate).

The poll has a margin of error (MOE) of ± 3%

This means…

The “true” percentage favoring proposal C is between 50% and 56%

Graphically we have:

Range of Percentages for Voters Favoring Ballot Proposal C

<---(-------|-------)----------------------------------->

50 53 56

Since a portion of the confidence interval is at 50%, the reporter <u>cannot</u> correctly claim that a majority of voters favor proposal C.

In other words, we conclude that there is <u>not</u> statistical evidence, at the 95% confidence level, that a majority of voters favor proposal C.

MORE ON CONFIDENCE INTERVALS AND SAMPELING ERRORS

Recall that confidence intervals are based on samples gathered from a larger population and when sampling from a population there is an error associated with the process of sampling. This error partially is based on the size of the sample. If we were to conduct a census (sampling all elements of a population) the sampling error would be zero.

There is an inverse relationship between sampling errors and the size of the sample. Accordingly, as the sample size decreases, the sampling error increases. In other words, the larger the sample size the smaller the error. Conversely, the smaller sample sizes yield larger the sampling errors.

Referencing the afore depicted polling example, the smaller sample of n = 400 has a 5% sampling error, Whereas, the larger sample of n = 1000 has a smaller sampling error of only 3%.

CONFIDENCE INTERVALS AND PROBABILITY

A common misperception is that a confidence interval is “directly” associated with the probability of the variable being measured. For example, consider the statement --- “There is 95% confidence that the true proportion of voters who favor candidate A is between 42% and 48%.” The 95% interval is <u>not</u> a range of probabilities for the proportion of voters who favor candidate A; rather, the 95% is associated with the confidence in the success rate of the procedure/process used to generate the interval.

Hence, it is not correct to say that there is a 95% probability that the true percentage favoring candidate A is between 42% and 48%. Similarly, it is not true that the interval has 95% probability to include the population parameter.

In actuality, the probability that the true proportion (i.e., the population parameter) falls within the confidence interval is either 0 or 1. In other words, a sample is used to generate an interval that either contains the population parameter (i.e., the true proportion), or it does not contain the population parameter; there is no way of being certain, short of sampling everyone. In other words, we cannot know the true population parameter without conducting a census.

If we were to select many different samples (of the same size) and construct the corresponding confidence intervals, 95% of them would actually contain the value of the population proportion and 5% of the time the true population proportion will lie outside the interval. This is a subtle, but important, difference from claiming that there is a 95% probability that the true parameter lies between the two values (bounds) of the interval.

ADDENDUM

Confidence intervals are associated with the realm of Frequentist Statistics**;** this is the most common school of statistical thought. However, there is a sphere of statistical reasoning known as Bayesian Statistics. In the Bayesian realm, there is a direct connection between probability and the interval of a parameter being estimated. These types of intervals are called "credible intervals".