UNDERSTANDING CONFIDENCE INTERVALS

POLLING RESULTS EXAMPLE

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.

A **sample survey** is a survey that is carried out using a sample of people who are intended to represent a larger population. For example, if we wanted to know more about the opinions of the one million adults who live in a particular city, and if we selected a sample of 900 individuals from those one million people, a survey conducted using those 900 people would be considered a sample survey. Sampling is used to estimate the values of population characteristics when a census of an entire population is not practical.

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 called a sample) and come up with an answer that is representative of what the population as a whole believes.

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?” Here is where statistics can use probability theory to quantify the results.

Accordingly, the results of a poll are reported in two parts… the **point estimate** and the **margin of error**.

For polling results, the point estimate is a single value that estimates the true population percentage associated with the poll question.

The margin of error is an interval estimate that accounts for the magnitude of the **sampling error** in the results of a survey. Typically, this estimate is express in terms of a plus and minus (±) percentage. Additionally, the margin of error has a **level of confidence** associated with the estimate; although, this confidence level value is rarely reported, it is typically 95%.
.

EXAMPLE 1

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

CANDIDATE A 54% **CANDIDATE B 46%** (these are point estimates)

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

This means…

The “true” percentage for **CANDIDATE A** is between 49% and 59%

The “true” percentage for **CANDIDATE B** is between 41% and 51%

Graphically we have:

CANDIDATE A vs. CANDIDATE B

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

<-----------(----------|----------)------------------------------------------->
            41         46         51
```

Since the confidence intervals overlap, we <u>cannot</u> claim there is a statistically significant difference (at a 95% confidence level) in the true percentage of voters supporting **CANDIDATE A** and the true percentage of voters supporting **CANDIDATE B**.

In other words, we conclude that there is not statistical evidence, at the 95% confidence level, that voters support **CANDIDATE A** over **CANDIDATE B**.

EXAMPLE 2

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

CANDIDATE A 54% **CANDIDATE B 46%** (these are point estimates)

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

This means…

The “true” percentage for **CANDIDATE A** is between 51% and 57%

The “true” percentage for **CANDIDATE B** is between 43% and 49%

Graphically we have:

CANDIDATE A vs. CANDIDATE B

```
<--------------------------------------------(-------|-------)------------------------>
                                             51      54      57

<-----------------------(-------|-------)---------------------------------------------->
                        43      46      49
```

Since the confidence intervals do not overlap, we can claim there is a statistically significant difference (at a 95% confidence level) between the true percentage of voters supporting **CANDIDATE A** and the true percentage of voters supporting **CANDIDATE B**.

In other words, we conclude that there is statistical evidence, at the 95% confidence level, that likely voters support **CANDIDATE A** over **CANDIDATE B**.

EXAMPLE 3

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

CANDIDATE A 53% **CANDIDATE B 47%** (these are point estimates)

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

This means…

The "true" percentage for **CANDIDATE A** is between 50% and 56%

The "true" percentage for **CANDIDATE B** is between 44% and 50%

Graphically we have:

CANDIDATE A vs. CANDIDATE B

```
<-----------------------------------------(-------|-------)----------------------->
                                          50      53      56

<------------------------(-------|-------)------------------------------------------->
                         44      47      50
```

Since the confidence intervals overlap, we cannot claim there is a statistically significant difference (at a 95% confidence level) in the true percentage of voters supporting **CANDIDATE A** and the true percentage of voters supporting **CANDIDATE B**.

In other words, we conclude that there is not statistical evidence, at the 95% confidence level, that voters support **CANDIDATE A** over **CANDIDATE B**.

CONFIDENCE INTERVALS AND SAMPELING ERRORS

Confidence intervals are based on samples gathered from a larger population. When sampling from a population there is an error associated with the process of sampling. This error 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. As the sample size decreases, the sampling error increases. Consequently, the larger the sample size the smaller the error. Conversely, the smaller the sample size the larger the error.

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 NOT 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 in the interval.

SIDE NOTE: Confidence intervals are associated with the realm of ***Frequentist Statistics*****;** 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”.