1 The Mathematics Behind Polling

1.1 Introduction

One place where we see inferential statistics used every day is in polling. Opinion polls, like it or not, are part of our political system. Polling is also used in marketing, sales, and entertainment. The intricacies of polling are far too complicated for us to treat completely. We can, however. understand the basic ideas behind this discipline and see how probability theory is used in polling. The view we take is greatly simplified and will necessarily gloss over some practical difficulties. It will, however, make it easier to interpret the kind of poll results normally reported in the news.

The basic idea is that in a large population of people a certain percentage will agree on one particular issue. We would like to know what percentage of people this is. Asking everyone and computing the exact result is out of the question given the size of the population. We might at least try to estimate the percentage by choosing a representative sample from the population and determining the percentage in the sample. Assuming that our sample is truly representative of the population, the percentage holding that opinion in the sample should provide a reasonable estimate of the percentage in the entire population.

Two natural questions might be, what exactly is "a reasonable estimate" and how confident are you in this estimate. Here is where statistics can use probability theory to quantify the results.

1.2 The Ideas Behind Statistical Estimation.

We will take a very simplified version of statistical estimation and see how the mathematics works. Remember, we are not going to deal with all the mathematical and practical issues that come with this method of estimation. We will just take the simplest possible version of estimation and explore the logic and mathematics behind it.

We start by assuming we have a very large population; so large that its actual size overwhelms any particular number we use to sample it. We would like to estimate the percentage of people in this population who would answer a specific question, "yes." For reasons that will become clear later, we restrict ourselves to questions where we expect a good number of people will give an answer of "yes" and a good number will give an answer "no." Let us denote this unknown percentage as $P_0\%$. We know it exists; it is just that we do not know what it is.

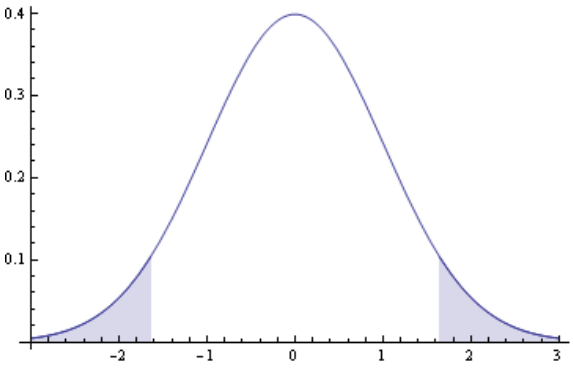
The first thing we do is consider the experiment of choosing one person at random from the population and asking them the question we are interested in. What is the probability that they will answer "yes"? If the selection process is actually random, than any one person is as equally likely to be chosen as any other. The people who will answer "yes" represent $P_0\%$ of that population.

That would mean that the probability that the person chosen will answer "yes" is $p_0 = \frac{P_0}{100}$. All we have done is convert the percentage to a ratio of the whole and changed that ratio into a real number between 0 and 1. Our one rule about our question is that a good number of people will answer of "yes" and a good number will answer "no." That means that p_0 should not be too close to 0 or to 1. Now that we are thinking of a random experiment where a particular event has probability p_0, we can consider that experiment as a Bernoulli trial where an answer of "yes" is a success. We still only know that the probability exists, and we do not know what it is.

Once we have chosen one person from the population, we do not what to choose them again. So we will not. Now our other overriding assumption about the population is that its actual size overwhelms any particular number of people we pick from it. This means that removing one single person from the population will have no measurable impact on the percentage who would answer "yes." If we choose a second person, the probability that they will answer "yes" is still p_0. The same goes for a third, fourth, or fifth. Thus we reasonably assume that every time we choose a person from the population, the probability they will answer "yes" is always p_0. That is to say, choosing n people randomly from the population amounts to repeating the same Bernoulli trial n times.

We know a lot about the probability model that comes from repeating a Bernoulli trial a number of times. We also have a quick way of computing probabilities in that model if the number of repetitions is large. Well we can if we actually know the probability of success in one trial. We do not just yet, but let us go on.

If we choose a large sample of people, say $n = 100$, $n = 500$, $n = 1000$ or more, the probability model should be very close to the normal distribution. Now as usual in a random experiment, anything can happen, but we still know what to expect. We expect that the number of successes in the sample will be close to the mean of the experiment, and that it be within one or two standard deviations of this mean. Unfortunately we do not know this mean or this standard deviation, but that does not change where the result would be if we did know them. We expect that the result will end up somewhere in the middle part of the normal distribution approximating the probability model



We have a chart that quantifies various choices for defining the center of the distribution:

Confidence	Lower limit	Upper limit
90%	-1.645 S.D.	1.645 S.D.
95%	-1.92 S.D.	1.92 S.D.
99%	-2.575 S.D.	2.575 S.D.

We can use this chart and say that we are 90% confident that the sample mean will be no more than 1.645 standard deviations away from the actual mean. We could claim 95% or 99% confidence by choosing no more than 1.92 or 2.575 standard deviations.

Here is where we draw a inference from a particular sample. We will go out a collect a random sample of 1000 people and ask them our question. We will find out exactly how many of these people answer "yes." Suppose that this turns out to be 675 of the 1000 sampled.

If the sample is truly random, this result should reflect the entire population quite closely. Thus we will consider the percentage of people in the sample who answer "yes" to be a close approximation to the percentage of people in the entire population who would have answered "yes" if asked. As a ratio to the whole, the ratio of people in the sample should be a close approximation to the percentage in the entire population. From the point of view of a Bernoulli trial, this ratio should be a good approximation of the probability p_0 of "success" in one trial.

Thus we will assume that estimate obtained from the sample $0.675 = \frac{675}{1000}$ is close enough to the actual probability p_0 to use it in its place. We can compute the expected mean, variance and standard deviation of the sampling experiment using the formulas:

$$\begin{aligned} \mu &= np_0 \\ \sigma^2 &= np_0(1-p_0) \\ \sigma &= \sqrt{np_0(1-p_0)} \end{aligned}$$

However, since we do not know p_0, we use the approximation

$$p_0 \simeq 0.675.$$

We the compute a sample mean

$$m = ns = 1000 \cdot \frac{675}{1000} = 675.$$

A sample variance

$$\begin{aligned} d^2 &= ns(1-s) \\ &= 1000 \cdot (0.675) \cdot (1-0.675) \\ &= 1000 \cdot (0.675) \cdot (0.325) \\ &= 219.38 \end{aligned}$$

And finally a sample standard deviation

$$d = \sqrt{ns(1-s)} = \sqrt{219.38} = 14.811.$$

Since we are 90% confident that the sample mean will be no more than 1.645 standard deviations away from the actual mean, we can say we are 90% confident that the actual mean will be no more than 1.645 sample standard deviations away from the sample mean So we are 90% confident that the actual mean μ is between

$$m - 1.645d \quad \text{and} \quad m - 1.645d.$$

That is to say we get

$$\begin{aligned} 675 - (1.645) \cdot (14.811) &\leq \mu \leq 675 + (1.645) \cdot (14.811) \\ 650.64 &\leq \mu \leq 699.36.. \end{aligned}$$

But we know the relationship between the population mean and the population probability, $\mu = np_0$. Thus

$$650.64 \leq 1000p_0 = \mu \leq 699.36.$$

So we are 90% confident that the actual probability p_0 is in the interval

$$0.65064 \leq p_0 \leq 0.69936.$$

Converting this to a percentage and doing a bit of rounding off, we have estimated, with 90% confidence that the percentage of people in the population who would answer yes to our question is between 65% and 70%. In other words, the percentage is approximately 65% with a margin of error of $\pm 2.5\%$ and a confidence level of 90%.

1.3 Examples

Example 1 *Suppose you would like to estimate the percentage of people in Arizona that say they enjoy the summer heat. You survey* 500 *people and find that* 267 *of them say that they do. This translates into a ratio of the whole of* $\frac{267}{500} = 0.534$ *or a percentage of* 53%. *What is the margin of error in the estimate if you use a confidence level of* 90%*?*

First, the sample size is $n = 500$. We represented the number of people who answered yes as a ratio of the whole: $s = 0.534$. This will approximate the unknown population ratio:

$$p_0 \simeq s = 0.534.$$

When we know the population probability, the formulas for the population parameters are

$$\begin{aligned} \mu &= np_0 \\ \sigma^2 &= np_0(1-p_0) \\ \sigma &= \sqrt{np_0(1-p_0)} \end{aligned}$$

We use these and the approximation of p_0 to compute a sample mean, a sample variance, and a sample standard deviation

$$\begin{aligned} m &= ns = 500 \cdot 0.534 = 267 \\ d^2 &= ns(1-s) = 500 \cdot 0.534 \cdot 0.466 = 122.82 \\ d &= \sqrt{ns(1-s)} = \sqrt{122.82} = 11.082. \end{aligned}$$

So we are 90% confident that the actual mean μ is between

$$m - 1.645d \quad \text{and} \quad m - 1.645d.$$

That is,

$$\begin{aligned} 267 - (1.645) \cdot (11.082) &\leq \mu \leq 267 + (1.645) \cdot (11.082) \\ 248.77 &\leq \mu \leq 285.23. \end{aligned}$$

But $\mu = np_0$. Thus

$$248.77 \leq 500p_0 \leq 285.23..$$

So we are 90% confident that the actual probability p_0 is in the interval

$$0.49754 = \frac{248.77}{500} \leq p_0 \leq \frac{285.23.}{500} = 0.570\,46.$$

Thus the population percentage P_0 is in the interval

$$49\% \leq P_0 \leq 58\%.$$

We round off making sure to widen the interval so that we do not loose any confidence in our estimation interval.

The final result we obtain is an estimate of 53.5% within an error of ±4.4% and a confidence of 90%.

Example 2 *Suppose we approximate the percentage of people over the age of 30 in Tucson by taking a random sample of people. It turns out that of* 5000 *people chosen* 3254 *are over 30. What result will we obtain using a* 90% *confidence level?*

The sample size is $n = 5000$. The sample gives a ratio of $s = \frac{3254}{5000} = 0.6508$. We use this to approximate the unknown population ratio

$$p_0 \simeq s = 0.6508.$$

in the formulas for the population parameters:

$$\begin{aligned} \mu &= np_0 \\ \sigma^2 &= np_0(1-p_0) \\ \sigma &= \sqrt{np_0(1-p_0)} \end{aligned}$$

That allows us to compute a sample mean, a sample variance, and a sample standard deviation

$$\begin{aligned} m &= ns = 5000 \cdot 0.6508 = 3254 \\ d^2 &= ns(1-s) = 5000 \cdot 0.6508 \cdot 0.3412 = 1110.3 \\ d &= \sqrt{ns(1-s)} = \sqrt{1110.3} = 33.321. \end{aligned}$$

As usual, a 90% confidence means that the actual mean μ is between

$$m - 1.645d \quad \text{and} \quad m - 1.645d.$$

That is,

$$\begin{aligned} 3254 - (1.645) \cdot (33.321) &\leq \mu \leq 3254 + (1.645) \cdot (33.321) \\ 3199.2 &\leq \mu \leq 3308.8. \end{aligned}$$

Using $\mu = np_0$,

$$3199.2 \leq 5000p_0 \leq 3308.8.$$

So we are 90% confident that the actual probability p_0 is in the interval

$$0.63984 = \frac{3199.2}{5000} \leq p_0 \leq \frac{3308.8}{5000} = 0.66176.$$

Thus the population percentage P_0 is in the interval

$$63\% \leq P_0 \leq 67\%.$$

We have found an estimate of 65% within an error of $\pm 2\%$ and a confidence level of 90%.

1.4 Final Comments

This gives a flavor of the statistical methods used in polling. Indeed if you look at news reports of polling results you will see data similar to what we produce. You will almost see a percentage estimate and bound on the error. Often you will see the number of people sampled reported as well. The confidence level is often left out of news reports, or even if it appears, it is buried in the article or in small type. This is probably because it is technical information that editors assume would confuse people not familiar with the language and methods of statistics.

It is important to realize that the practical practice of polling, especially political polling, is much more sophisticated than the examples above let on. One large, and clearly major issue in polling is choosing a representative random sample. As we have seen, random means unpredictable. Thus a process that controls the selection of people to be sampled should not single out particular groups, locations, or ages to be truly random. In polling a random sample must also be representative. If a political pollster does not limit his polling to

citizens of a legal age to vote, the resulting sample is not likely to be representative of the population he wants to study. As a result as much statistics goes into the process of selecting a random sample as in analyzing the data collected from the sample.

A final note is that we have only seen two simple examples of statistics in use. There are many, many more. There are methods, not that dissimilar from the ones above that apply when the problem limits the size of a sample or test. These use other mathematical distributions other than the normal distribution. There are also estimation methods that can be used to sharpen the results of a complete survey of a large population. Rather than beginning with a small sample and roughly estimating counts in an entire population, these techniques take the results of a comprehensive survey of a population and adjust the raw data to account for errors in the counting and tabulation of the data collected. The US census bureau uses these techniques to strengthen the quality of the results they report involving demographic information about the country. However, there is a long-standing controversy that does not allow them to use these results on the data used by congress to apportion legislative representation or the distribution of federal funds to states and local communities. While this debate often revolves about the veracity of the statistical methods, the true issue is the perceived advantages that using statistics or not using statistics might have for one party or the other.

Prepared by: Daniel Madden and Alyssa Keri: May 2009