Inferential Statistics

Table of Contents

INTRODUCTION

Inferential statistics allow us to make predictions about the total population based on a sample of the population. In other words, inferential statistics allows us to generalize the results we have collected from a sample of individuals to the larger population. In many practical cases, the population is what we really want information about, but the sample is what we can get.

KEY POINTS IN THIS DOCUMENT

- It is not always possible to measure an entire population. Inferential statistics allow us to make predictions about the total population based on characteristics of a sample of the population.
- **Sampling error** is the error introduced by not surveying the entire population.
- **The standard error of the mean** is the amount of uncertainty there is about the population mean. It is a measure of how good or bad our estimate is.
- **Confidence intervals** are ranges that we can say some unknown value (e.g., the population mean) is within some pre-set level of confidence.

SAMPLING ERROR

The **sample** is the subset of the population that is selected to respond to a survey. If you would like to be able to project/generalize the results of a sample survey onto your population of interest, you *must* quantify the **sampling error** of the survey.

Sampling error is the error introduced by not surveying the entire population. It refers to the difference between estimates based on your sample population and the actual reality in an entire population of interest. Sampling error can be controlled by sample size and sampling methodology.

Inferential statistics allow us to estimate how closely the characteristics of the sample reflect the characteristics of the population, based on the sampling error. In order to use the sample to infer the characteristics of the population, appropriate sampling approaches, such as probability sampling, are required.

STANDARD DEVIATION

From a sample, we can measure some statistics exactly (e.g., sample mean and sample standard deviation) and we can often estimate some parameters (e.g., the population mean and population standard deviation). The population standard deviation is mainly useful in revealing how good a particular guess or inference about the population mean is.

As an example, imagine you are interested in testing the weight of peanut butter in jars. Each jar claims it has 1000 grams, but the machines that fill the jars are not perfect, so the weight per jar varies a little bit.

After scooping out the peanut butter from five (n=5) jars, you measure x = {1002, 1004, 997, 998, 1001} grams. The sample mean is x-bar = (1002 + 1004 + 997 + 998 + 1001) / 5 = 5002 / 5 = 1000.4 grams.

The population mean, however, we do not know, and we cannot know it without measuring every jar that comes out of the factory. We can make an educated guess though based on the assumption that the sample of 5 jars that we did measure came from a random sample of all the jars. If that is the case, then it is fair to infer that the population mean is also 1000.4. **The difference here is that we know the sample mean, but we're estimating the population mean**.

How good is this estimate? We can find out with the sample standard deviation, which is s = 2.881 (see the downloadable document "Descriptive Statistics" for an explanation of standard deviation). We'll assume from this statistic that the population standard deviation is the same, sigma =2.881.

STANDARD ERROR

The standard error of the mean is the amount of uncertainty there is about the population mean. It is a measure of how good or bad our estimate is. The bigger the standard deviation is, the worse our estimate becomes (and the bigger the standard error becomes). However, the more data in our sample, the better our estimate becomes (and the smaller the standard error).

It takes four times as many observations to make an estimate twice as good.

The formula for standard error of the mean is s / sqrt(n), so with our peanut butter jars guess,

standard error = 2.881 / sqrt(5) = 1.288 grams.

Our estimate of the population mean was 1000.4 grams, but typically, an estimate like that is going to be wrong by 1.288 grams. We don't know which way the guess will be wrong, but it's going to be wrong by at least a little most of the time, and that is okay.

CONFIDENCE INTERVALS

With an estimate of the population mean and the standard error of that mean, we can calculate a confidence interval. **Confidence Intervals** are ranges that we can say some unknown value (in the example given, the population mean) is within some pre-set level of confidence.

In the peanut butter jar example we can be…

- 80% sure (or confident) that the population mean is between 998.43 and 1002.37 grams,
- 90% sure that the population mean is between 997.65 and 1003.15 grams,
- 95% sure that the population mean is between 996.82 and 1003.98 grams, and
- 99% sure that the population mean is between 994.47 and 1006.33 grams.

The exact calculations for these confidence intervals are not very important because they are usually done with a computer in practice. There are a few features and patterns that are important.

- Every confidence interval should have a a) lower bound, b) an upper bound, and c) a confidence level (e.g., 996.82 to 1003.98 with 95% confidence).
- If you ever see a confidence interval without the level of confidence reported, the confidence level is likely 95%.
- The estimate (1000.4 grams) is always within the confidence interval. For a mean estimate like the one in this example, it is right in the middle.
- In order to have greater certainty (that is, to have higher confidence) without collecting more data, the interval needs to become wider.
- The only way to get 100% confidence is to make the interval so wide that it is not useful (the mean weight per jar is between 0 and 999 999 grams with 100% confidence)
- Sometimes estimates and confidence intervals are reported together as "(estimate) +/- (margin of error), 95% of the time". "+/-" , or "plus-or-minus" just means the estimate could be above or below the true value, and "margin of error" is the amount that estimate could be wrong by.