2

Point Estimation

Recall. One of the most important goals of statistics is to make inferences about population parameters based on samples of data.

The most basic and common type of inference is to compute a single-numbered statistic and use it to estimate a parameter.

Example 2.1. To estimate a population mean μ, you could collect a random sample of data $y_1, y_2, \ldots, y_n$ from the population and compute the *median*. That is then your estimate of μ. Alternatively, you could compute the *sample mean*

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

and use that as your estimate of μ. This leads to a natural question. Which estimate is better, the median or the mean? The answer is that it really depends on both the survey sample design AND on your concept of "better." In fact, you were probably told in Stat 160 that the median is a better measure of centre when you have a skewed population distribution, whereas the mean is a better measure when your population distribution is roughly symmetric and unimodal. Although this is adequate for a first-year class, it is time to give a more sophisticated answer.

Random variables and realizations

Statisticians often use take *dual* view of random variables by viewing both a random variable and its realization as the same thing. Although this is extremely convenient, it is not quite precise, and often leads to confusion for students when first encountered.

This dual view can be explained as follows. BEFORE the experiment, we have some random variables $Y_1, Y_2, \ldots, Y_n$ which will become the data once they are observed.

Often, but not always, it is reasonable to assume that

- $Y_1, \ldots, Y_n$ are *independent*,
- $Y_1, \ldots, Y_n$ all have the same distribution function (i.e., they are *identical in distribution*),
- the distribution belongs to some nice family (such as the normal, Poisson, exponential, binomial, etc.), and
- the value of the population parameter of interest is fixed, but unknown.

Notation. We say that $Y_1, \ldots, Y_n$ are *i.i.d.* if they are independent and identically distributed. We use the phrase *random sample* synonymously with i.i.d.; that is, $Y_1, \ldots, Y_n$ are a random sample if and only if $Y_1, \ldots, Y_n$ are i.i.d.

AFTER the experiment, we have observed data $y_1, y_2, \ldots, y_n$ (i.e., *realizations* of $Y_1, Y_2, \ldots, Y_n$).

Notation. It is important to mention that we will use the phrase *random sample* in this dual sense. We say that $y_1, \ldots, y_n$ are a *random sample* if $y_1, \ldots, y_n$ are the data that result as realizations of the random sample $Y_1, \ldots, Y_n$.

The goal of *estimation* is to use the data $y_1, y_2, \ldots, y_n$ to construct a best guess or *estimate* for the parameter value of interest.

Notation. As we have already mentioned, a *statistic* is a number computed from data. However, this use of statistic is too restrictive. Instead, we will use the word *statistic* in the dual sense; that is, (i) if $Y_1, \ldots, Y_n$ are a random sample and $g(Y_1, \ldots, Y_n)$ is a function of $Y_1, \ldots, Y_n$, then we say that $g(Y_1, \ldots, Y_n)$ is a *statistic*, and (ii) if $y_1, \ldots, y_n$ are a random sample of data and $g(y_1, \ldots, y_n)$ is a function of $y_1, \ldots, y_n$, then we say that $g(y_1, \ldots, y_n)$ is a *statistic*. It is worth mentioning that case (ii) really does say that a statistic is a number computed from data. Since $y_1, \ldots, y_n$ are numbers, $g(y_1, \ldots, y_n)$ is just a number.

If this dual use of statistic (a random variable on the one hand and a number on the other) is confusing, then you will be relieved to know that in *estimation* there are distinct words used. An *estimator* is a random variable, whereas an *estimate* is a number.

Formally, suppose that θ is a population parameter and that the estimation of θ is desired. For example, θ could be the population mean (traditionally called μ) or the population variance (traditionally called σ^2), or some other parameter (such as the population maximum, the population minimum, the population median, etc.). Let $Y_1, \ldots, Y_n$ be a random sample from this population. In particular, $Y_1, \ldots, Y_n$ are i.i.d. from a distribution that depends on the parameter θ. An *estimator* of θ is a statistic $g(Y_1, \ldots, Y_n)$ whereas an *estimate* of θ is a statistic $g(y_1, \ldots, y_n)$.

Traditionally, we use the notation $\hat{\theta} := g(Y_1, \ldots, Y_n)$ for an estimator of θ.