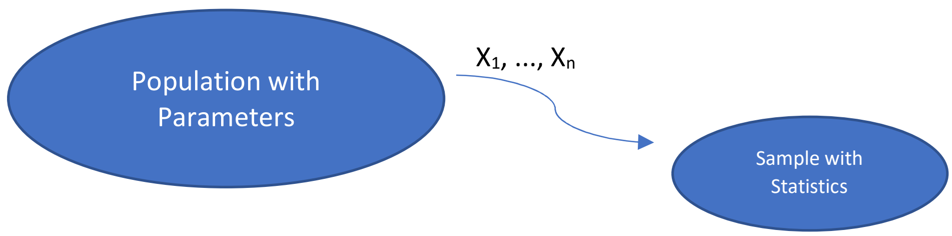
Confidence Intervals and Hypothesis Testing
The Big Picture

Confidence intervals and hypothesis tests reflect two kinds of statistical inference: ***estimation*** and, well, ***hypothesis testing***. Before getting a feel for these two types and how they relate to each other, it will help to look at a typical setting for statistical inference.

Typical Setting for Inference



We have some population, say all the eligible and registered voters for an upcoming election to select a leader. We can associate what we call ***parameters*** with this population.

> For example, a parameter might be the *proportion* of eligible and registered voters who, on October 23, 2020, intended to vote for Joseph Biden in the upcoming US presidential election.

For now, let's treat these parameters as *fixed quantities*. That is, we do not treat them as *random*: the proportion of voters on October 23, 2020 intending to vote for Joseph Biden is taken to be fixed. We don't know (exactly) what this proportion is, so we treat it as *unknown*, but that doesn't make it random and therefore subject to change. It is a fixed, unknown quantity.

For reasons which are usually practical (perhaps it is too expensive to perform all the required measurements or calculations), we usually cannot observe or measure, directly or indirectly, what these parameters are. So we *infer* what they are by drawing on what we can observe, measure, or calculate. From a much smaller subset ($X_1, \ldots, X_n$) of the overall population we calculate what are called ***statistics***. We call the subset a *sample* and so long as this sample is representative of the larger population so that its own characteristics mirror those of the population, we can infer what the characteristics of the population are from those of the sample. That is, when we calculate the right statistics, we can infer what the parameters of the population are from the statistics we calculate from the sample.

The primary way to ensure that any sample we draw from the population is representative of this population is to draw a *random sample* (there are various ways to do this, of course). For this reason the statistics calculated from the samples, unlike the parameters, *are* random (what we call *random* *variables*) because they are calculated from measurements on elements of random samples. And because they are random variables, there are probability distributions associated with them. These probability distributions, if we know them at least approximately, are extremely important: we don't use statistics alone but also their associated distributions to infer what the population parameters are.

> For example: on October 23, 2020, we might conduct a (random) survey of 118 eligible and registered voters for the upcoming presidential election and calculate the proportion of these voters intending to vote for Biden. This sample proportion is a statistic and it has an associated probability distribution. We use the sample proportion along with the distribution in statistical inference.

<><><><><><><>

Now to two kinds of statistical inference: ***estimation***, which involves confidence intervals, and ***hypothesis testing***. These two kinds complement each other and each attempts to answer a distinct question about the population parameter of interest.

Estimation

The guiding question for estimation is: what is the value of the population parameter?

To answer this question, we use sample statistics as *point estimates* of the value of a population parameter. These point estimates most likely do not equal the exact value of the population parameter but, if we choose the right statistic and sample appropriately so that the sample is representative of the population, should be "close" to it.

But how close is "close"? That's where probability distributions come in. Together with either exact or (most often) approximate probability distributions for our sample statistics or our point estimates, we construct intervals out of these point estimates. We call these intervals *confidence intervals* because, given how we construct them, we attach probabilities or *confidences* to them. For example, a 95% confidence interval might be:

> There is a 95% probability that the interval (48%, 53%) contains the percentage or proportion of eligible and registered voters intending to vote for Biden in an upcoming election.

Hypothesis testing

The guiding question for hypothesis testing is: how likely is it that a population parameter is some fixed value or some value in a set of fixed values?

With hypothesis testing, we are not so much interested in estimating a parameter as in testing whether it has a specific value or value within a range of values. To conduct a test, we first set up hypotheses about a parameter value, and then use sample statistics together with their probability distributions to test these hypotheses.

> One might hypothesize that no more than 50% of eligible and registered voters intend to vote for Biden in an upcoming election and then use a sample proportion together with its probability distribution to reject or fail to reject this hypothesis.

How do confidence intervals and hypothesis tests complement each other?

While a confidence interval gives us some idea of where the value of a parameter lies, hypothesis tests help quantify the amount of evidence for or against the parameter having some value or set of values. Together, they help us learn about the underlying population parameters.

> Suppose we have two different confidence intervals (from two different random samples) for the proportion of eligible and registered voters intending to vote for Joseph Biden: the first is (51%, 56%) and the second is (53%, 58%). Using these intervals together gives us a better idea of where the value of the parameter lies than either one alone. In addition, note that the second one provides more evidence against the hypothesis that no more than 50% of voters intend to vote for Biden than the first one. But how much evidence? A hypothesis test helps us quantify this amount by providing a metric to measure it.