**Definitions and terms to know – Part 4: Confidence intervals**

**Population parameter** – these are typically unknown, fixed numbers representing the population proportion, mean, variance, and standard deviation. There are many other types of parameters but these are the ones we are most interested in for now.

* For categorical data
  + Population proportion:
* For quantitative data
  + Population mean:
  + Population variance:
  + Population standard deviation:

**Sample estimate** – these are values that approximate the population parameters, we can view them as random variables and consider their sampling distributions, or, once we observe a data set, we can plug in the data and get a point estimate which is an actual numerical guess

* For binary, categorical data
  + Sample proportion:
* For quantitative data
  + Sample mean:
  + Sample variance:
  + Sample standard deviation:

**Sampling distribution** – the theoretical distribution of any sample estimate, typically if our sample size is large enough, we can use the Central Limit Theorem to approximate this theoretical distribution with a Normal model

* Theoretical sampling mean: usually this is the true unknown value of the population parameter (e.g. or )
* Theoretical sampling variance (called sampling error): is function of the true unknown population parameter and the sample size (e.g. or )
  + Estimated standard error (called the **standard error** and abbreviated ) is the estimated standard deviation of a sampling distribution (e.g. or )

**General formula for any confidence interval:**

* The sample estimate is always in the middle of the CI
* **Margin of error**:
* The **critical value** is a quantile from a particular distribution (most often from a standard normal distribution) that matches with some *significance level* (see below).

**Confidence interval assumption:** The data was collected without bias and each observation is independent of the others and we can apply the CLT.

**When to use CIs:** When the population parameters are unknown and we want a range of possible values for these unknown parameters.

**Honest interpretation of a** **CI:**

(Note: Anywhere you see <>’s you should replace the inside with problem-specific terms.)

“We don’t know exactly what <the unknown parameter value> is, but the interval from <lower bound> to <upper bound> probably contains the true <parameter value>.”

OR

“We are confident that the true population <parameter> is between <lower bound> and <upper bound>. “

The key to these interpretations is that we admit that we are unsure about two things. First, we need an interval, not just a single value, to try to capture the true <parameter value>. Second, we aren’t even certain that the true <parameter value> is included in our interval, but we’re “pretty sure” that it is. By pretty sure, we mean in the hypothetical long-run frequency sense. We are saying that if we could take all random samples of size from this population and then create a CI based on each of these different random samples, then of these confidence intervals will contain the true population <parameter value>.

is called the **significance level** and is called the **confidence level**.

**Tips on how to interpret CIs**

Confidence intervals are useful because they allow us to quantify our uncertainty. We can heuristically think about them as informing us on how much we'd be willing to bet on certain outcomes. What we really mean however, has to do with the behavior of these intervals if we could calculate them based for *all possible samples* of size from the population of interest.

* **Don’t** 
  + Claim that other samples will agree with yours. A single CI makes a statement about the true (unknown) population proportion based on a single sample of data.
  + Ever state absolute certainty about the true value of a parameter. In statistics (regardless of the probabilistic framework), when we say we know something with certainty, this means that it occurs “with probability 1”.
  + Draw conclusions beyond the scope of the population from which your sample was collected. Be very careful about defining the population for which your sample is representative.
  + Suggest that the true parameter value changes. A confidence interval varies from one random sample to the next. The parameter stays the same (in frequentist statistics).
* **Do**
  + Treat the entire range of numbers within the confidence interval equally. Values near the middle are not "more plausible" than values near the edges.
  + Beware of margins of error that are too large to be useful. There is always going to be a trade-off between how precise our interval is vs. how confident we are in our interval.
  + Watch out for biased sampling. If the data you are using to build the CI is biased, then your CI will also be biased, thus making the quantification of our uncertainty less accurate.
  + Always think about ways in which the independence assumption might be violated. We usually can't check this assumption but if it is violated, it makes our uncertainty quantification less accurate.