CENTRAL LIMIT THEOREM

Of course a large sample is better than a small one - or even a bunch of small ones. The issue is that we seek to know the "truth" about the population. Such information isn't merely unknown - it's unknowable. That is, we cannot directly find the mean or the proportion or the standard deviation or the precise distribution for a population. What's the average normal human body temperature? What proportion of people have high blood pressure? How large a risk factor for colon cancer is a high-fat diet? Such questions are understandable and important, but can't be definitively answered.

The best we can do is carefully collect some data from as large and representative a group of individuals as possible and then start guessing. We evaluate a statistic for our sample. We think it gives us a good guess as to the value of the population parameter. But we know it's only a guess, so we also want to know how close we might be to the truth. Anyone can say that their sample results indicate that the population mean is close to such-and-such. It's the next step that matters: being able to say that that the true value we wish we knew is almost certainly within a certain distance above or below the one we calculated.

We don't actually collect 50 samples, then. We collect one. We merely (and powerfully) imagine how others might behave. We know that different samples would produce different estimates. We see, then, that sampling error is unavoidable. The truly amazing thing is that it's predictable and quantifiable! That's the magic of the CLT. It tells us that regardless of the unknowable features of the population, this sampling error follows (miracle!) a Normal model, and it further tells us how much variation to expect (the SD of that sample-to-sample variability). From that stunning knowledge we can reason backwards from our sample to make educated guesses about the things we wish we knew.

Look at it this way. In Statistics we usually aren't happy with what we know. We wish we knew something *else*, something grander. We want to expand the limited information we observe into insights about the workings of the large and complex world. If the CLT weren't true, we'd be done for. We'd collect our data and say that in our sample such-and-such happened. We'd say we know that this result isn't precisely true about the population. And then we'd be done. We'd have no idea how to elevate our insight to say things like "the population value must almost surely be within 2% of this" or "if the population behaved the way I thought it did what I see here almost surely could not have happened, so I must have been wrong".

As I tell my students, absent the CLT, the rest of Statistics does not exist. AP Stats would end in January -- but no such luck! In fact, we're far luckier. The cool stuff - inference - lies ahead. It's how we discover the truth about the world. What could be more exciting or empowering?

- Dave Bock