Sample Blog

Author Name

2019-01-19

Contents

Chapter 1

Introduction

Brief motivation behind the blog

Chapter 2

The magic of the central limit theorem:

2.1 Sampling, sampling, sampling...

As scientists we aim to understand the world around us, not just our immediate environments. In most cases, we don't have access to **populations**, for one, because they are... large. For example, if you're studying expectant mothers, it is virtually impossible to collect data from all of the expectant mothers from around the world. Therefore, we make do with random and representative *samples* of the population to make generalizations – that is, statements – about the population as a whole.

The **central limit theorem (CLT)** states that the larger the sample size collected, the closer the distribution of the *sample* means will resemble a normal distribution (*bell curve*), regardless of the population's distribution. *If you were asleep during your stats class or need a refresher, Khan Academy gives a good introduction to CLT.*

It's useful to re-read the statement above, because we're not saying that we're assuming that the observations (e.g., one sample of with 100) originate from a normal distribution. We're saying that the *distribution of sample means* (based on taking the mean of many separate samples that originate from the "parent" distribution) will be normally distributed. That is, provided your sample size is "large enough." The theorem works "in the limit", as mathematicians say, but a general rule is that the sample size should be at least 30 for the CLT to hold for almost any data distribution. And, in practice, it will work on smaller samples if they originate from a population that actually has a normal distribution.

Let's refresh ourselves with a few terms. **Variance** refers to the amount of variability, or how spread the data are from the mean. The population variance is symbolized as **sigma squared**, or σ^2. Because variance is a squared term, we tend to look at the square root of variance, or **standard deviation**, and the population standard deviation is symbolized as σ, or **sigma**.

We know that as the size of the sample increases, the closer the *sampling distribution of the sample mean will resemble a bell curve with a mean* that approaches the population mean, μ. How about the standard deviation of the distribution of the sample means? As the sample size increases, the CLT says that the standard deviation will approach $\frac{\sigma}{\sqrt{n}}$. Again, this results holds despite the shape of the population distribution.

A good source of intuitive discussion on the central limit theorem is Mordkoff, J.T. (2016) The Assumption(s) of Normality.

Post originally by Kelly Morrow and edited by L. Pessoa

Chapter 3

Applying the Central Limit Theorem

3.1 An example

According the National Center for Health Statistics, the distribution of serum cholesterol levels for 20 to 74-year-old males living in the United States has mean 211 mg/dl, and a standard deviation of 46 mg/dl. We are planning to collect a sample of 25 individuals and measure their cholesterol levels.

We are interested in the following about the sample:

1. What is the probability that our sample mean will be above a certain limit, say 230?
2. What is the 95% confidence interval of our sample means?
3. How do these vary as we collect more samples? Does the probability increase or decrease? Does the size of the confidence interval increase or decrease? By what factor does it increase or decrease?
4. Finally, how large would the sample size have to be to ensure a 95% probability that the sample average is within 5 mg/dl of the population mean?

How does the Central Limit Theorem (CLT) help us answer these questions?

3.2 Central Limit Theorem

Given a population with a finite mean μ and a finite non-zero standard deviation σ, the distribution of the sample means approaches a normal distribution as the sample size increases.

The mean of the sample means is given by

$$\mu_{\bar{X}} = \mu,$$

and the standard deviation of the sample means (also referred to as the standard error of the mean) is given by

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}.$$

For a more precise version of CLT, please refer Wolfram.

An important observation is that no assumptions are made about the distribution of the parental population. It could be discrete or continuous, severely skewed, but as long as the mean and standard deviation are finite, CLT holds. To convince yourself of this, please take a look at examples here or use the simulator here.

3.3 Samples

But what is a **sample**? We keep using that word, and its meaning is quite an important concept. A sample is a random draw of size N of data from the parent distribution. We obtain a sample of data every time we randomly draw from what we conceptualize as the population of interest. If the population of interest in every student at UMD, then a random draw is obtained by any mechanism that (truly) randomly draws from it (think of an imaginary lottery machine that, after spinning it, we can obtain one student at a time).

And the **sample mean**? The sample mean is just the average of the measure of interest from the N units that were sampled. So for each mean, we get exactly one sample mean.

When we're thinking about the CLT (what it means), we need to think of repeating this process many times to have *multiple sample means*. Remember, that each sample has N units. So for each mean, you need to sample a group of size N. This is what the CLT talks about.

One reason this might appear confusing is that in any one given study, we only sample once (with size N). That's the world the experimenter lives in (except it she repeats her experiment). But that's not the world of the CLT, which instead is a world in which we perform an experiment (a sample draw of size N), over and over. And a few more times…

Back to the previous questions. To answer them, it is essential to know the *sampling distribution*, that is, the *distribution* of the sample mean.

- Since the standard deviation of the parent population is known, from CLT it follows that the sampling distribution ($N = 25$) has a mean $\mu_{\bar{X}} = 211$ mg/dl, and standard deviation (this is called standard error) $\sigma_{\bar{X}} = \frac{46}{\sqrt{25}} = 9.2$. Note that the standard error reduces as the number of samples increase by a factor $\sqrt{N}$.
- Our limit, 230 mg/dl, is therefore $\frac{230-211}{9.2} = 2.07$ standard deviations away from the mean. In other words, the z-score associated with the limit 230 mg/dl is 2.07.

1. The answer to our first question is simple the probability that a normally distributed random variable is greater than 2.07 standard deviations away from the mean = 1.9% (or 0.019).
2. For a normally distributed random variable, 95% of the values lie within 1.96 standard deviations of its mean (on either side). The standard deviation remains the same, 9.2. Thus, the 95% confidence interval is simply, 211 - 1.96(9.2) = 193.0 to 211 + 1.96(9.2) = 229.0.
3. Suppose we had only 10 samples. Verify that the new standard error would be 14.5 and the z-score associated with the limit, 230 mg/dl, would be 1.31. The probability of our sample mean being over 230 would thus be 9.6% (almost 5 times higher). On the other hand, our confidence interval would be much larger with 10 samples; $(182.5, 239.5)$. How about if we had 50 samples? This would result is a narrower confidence interval $(198.2, 223.8)$ since the standard error is smaller.
4. To answer our final question, we need 1.96 standard deviations of the sampling distribution to amount to 5 mg/dl. Thus, the standard error should be $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = 5/1.96$. Since, we know σ, $N = 325.1$. We would need at least 326 samples to ensure this confidence interval.

3.4 Hypothesis Testing

Thinking along these lines can be used to develop hypothesis tests and understand p-values. We start again with an example:

Cystic fibrosis is a genetic disease that affects lung function. Forced vital capacity (FVC) is the volume of air that a person can expel from the lungs in 6 seconds. It is often used as a marker of the progression of cystic fibrosis. 14 participants received both a drug and placebo (at different times), and their FVC was measured at the beginning and end of each treatment period. In the study, the mean difference in reduction in FVC (placebo - drug) was 137, with a sample standard deviation 223. Does the drug have a significant effect?

The null hypothesis is that the drug has no effect, thus the reduction in FVC should be zero. Let's first find the probability of observing an FVC reduction of greater than 137 given the null hypothesis.

We calculate the standard error based on the 14 participants as $\sigma_{\bar{X}} = \frac{223}{\sqrt{14}} = 60$. The z-score associated with the mean reduction in FVC is given by $\frac{(137-0)}{60} = 2.28$. The probability of observing a value further from the mean by at least 2.28 standard deviations, which is also the p-value, which is 2.2% (or 0.022).

This is a small probability that the drug is having an effect just by chance. Why did we obtain a small probability? Because it's effect by chance should be zero. But because we're working with a sample ($N = 14$) that is randomly drawn from the population, the observed mean reduction will fluctuate from sample to sample. Based on data from our sample, the reduction was 137. But how large is 137? We don't know without some form of calibration, which can be obtained by the information that was given to us: $\sigma_{\bar{X}}$.

From these data, we can believe that the drug helps prevent deterioration in lung function. At least the data are consistent with this notion in a probabilistic sense. Another way to think about this, it would be somewhat irrational to think that we could obtain the observed result just by chance. Maybe not entirely with a p of basically $\frac{2}{100}$ but probably with a p of $\frac{1}{1000}$. But obviously this is somewhat subjective.

3.5 The flaw in the z-test

Is the above reasoning correct? To understand this we need to understand the difference between the first and second examples.

In our first example, an oracle provided us with the standard deviation of the population. Some all-knowing being told us what the population σ was. But in the second example the standard error was based on the sample. To make the distinction clear, we call the standard error based on sample data $s_{\bar{X}}$.

Important aside: why use the sample and not the population? Populations are essentially Platonic objects. We typically don't have complete knowledge about the objects we want to study. If we did we woulnd't need to study them in the first place! So we have to draw samples and do the best based on finite amounts of data. Another way to think about this is that oracles don't walk around waiting to be interrogated. Maybe they were around in ancient Greece, but not anymore…

Gossett (which published under the pseudonym Student) showed that when the standard error is estimated from sample data, the statistic $\frac{\bar{x}-\mu}{\sigma_{\bar{X}}}$ is not normal, but follows a t distribution with $N-1$ degrees of freedom (df). The t distribution looks very much like a normal, but has what we call heavier tails, that is, more mass along the tails relative to the normal.

Thus, the probability associated with a t-score of 2.28 with $14-1$ degrees of freedom can be calculated to be 4% (or 0.04). The p-value obtained from the z-test (2.2%) overstates the evidence against the null hypothesis (this is consistent with the fact that a normal is thinner along the tails than the t distribution).

Examples are borrowed from Introduction to Biostatistics kindly offered by Patrick Breheny at UIowa.

Post by Manasij Venkatesh, with edits by L. Pessoa.