1 Central Limit Theorem

1.1 Applications

Recall that you do not need to know the distributions of the X_i. You also – in many cases – do not need a very large sample size to obtain fairly accurate results. This is because, for many distributions (as long as the unknown distribution is not highly asymmetric or unusual), convergence to the Normal is often reasonably fast. So, generally speaking, $n \geq 30$ is a good sample size.

(Example.) In his 2020 interview with PBS NewsHour about his work on post-election auditing, the UC Berkeley statistician Professor Philip Stark gave the analogy of cooking a pot of soup.

In order to know if it is tasty/too salty/etc., you do not need to drink the whole pot of soup. Instead, as long as you mix up the pot sufficiently well, even just one spoonful is enough, no matter how large the pot is.

(Example.) A surveyor wants to measure the distance d between two locations A and B. She knows there will be some degree of error (due to human error, atmospheric distortions, etc.) Therefore, instead of taking just one reading, she decides to take $n = 36$ of them. Assuming the measurements are IID, and that the SD associated with measurements is $\sigma = 0.001$ (perhaps based on past experience), what can we say about the true distance d?

It is natural to expect that the expected value of any given reading is $\mu = d$ (at least, we would hope so). Of course, this is just an expectation. Now, if $X_1, X_2, \ldots$ are an IID sequence of measurements, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \mapsto d$$

as $n \mapsto \infty$. But, for just $n = 36$ measurements, what can we say about d? Can we at least estimate d within some reasonable degree of freedom?

Now, note that by CLT,

$$\frac{A_n - d}{\sigma/\sqrt{n}}$$

is asymptotically Normal(0, 1), where recall

$$A_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the sample average.

Now, if Z is Normal(0, 1), then $\mathbb{P}(|Z| \leq 1.96) \approx 95\%$ (here, we just arbitrarily picked 95% and then found 1.96 through an online calculator). Therefore,

$$\mathbb{P}\left(\left|\frac{A_n - d}{\sigma/\sqrt{n}}\right| \leq 1.96\right) \approx 95\%.$$

This is useful because everything here is known (we know the standard deviation, sample average, and n) *except* for the true distance d. Now, note that

$$\left|\frac{A_n - d}{\sigma/\sqrt{n}}\right| = \left|\frac{A_n - d}{0.001/\sqrt{36}}\right| \leq 1.96$$

if $d \in \left[A_n \pm 1.96\frac{0.001}{6}\right] = [A_n \pm 0.000327]$. This is known as a **confidence interval**.

Now, suppose that the sample average after 36 readings is 1.0045. Then, we would say that we are "95% confident" that the true distance d is somewhere in the interval $[1.0045 \pm 0.000327] = [1.004173, 1.004827]$. This is what is called a **95% confience interval (CI)**.

Remarks:

- Note that the sample average is denoted by $\overline{\mu}$, $\hat{\mu}$, $\overline{x}$, etc.
- Note that $[a \pm b] = [a - b, a + b]$.

More generally, to make a $(100)(1 - \alpha)\%$ CI for an unknown mean μ, supposing the true standard deviation σ is known, we

(a) Find z_* such that $\mathbb{P}(|Z| \leq z_*) = 100(1 - \alpha)\%$. Note that, for $\alpha = 0.1, 0.05, 0.01$ (corresponding to $90, 95, 99\%$ confidence), we have $z_* \approx 1.64, 1.96, 2.58$. As α decreases, naturally z_* (and hence the width of the CI) decreases.

(b) Find the sample average $\hat{\mu}$.

(c) Then, we can construct the confidence interval

$$\left[\hat{\mu} \pm z_* \frac{\sigma}{\sqrt{n}}\right].$$

Notice that there is a tradeoff:

- The width of the interval will increase as the confidence level increases.
- As we increase the size of the sample, the width of the confidence interval will decrease.

Interpretating what, for example, "95% confident" means here is theoretically subtle. For instance, notice that

$$\mathbb{P}\left(\mu \in \left[\hat{\mu} \pm z_* \frac{\sigma}{\sqrt{n}}\right]\right) = 95\%$$

is non-sensical. Just because we do not know μ dies not make it random; it either is in the CI or it isn't. Therefore, this probability is in fact either 0 or 1, but we do not know which.

Note that the confidence interval is random, not μ. This is because we took a random sample. Interpreting what "95% confident" means actually involves another application of the CLT.

Specifically, what we mean here is that if we were to build a large number of IID CIs, in exactly the same way that we did this *one* CI, then

- provided that n is reasonably large, about 95% of them would contain the true value of μ,
- and, in this sense, we are "95% confident" that this *one* CI we have made contains the true value of μ.

1.1.1 t-Distribution

More often in practice, both μ and σ are unknown. In this case, experience has shown that instead of using the Normal distribution, it is better to use an alternative − but related − distribution called **Student's *t*-distribution.** This makes the most difference when the sample size n is small. In fact, as $n \mapsto \infty$, the student's t-distribution converges to the standard Normal. So, if n is large, the improvement is negligible.

If the true standard deviation σ is unknown, then we must estimate it using the sample standard deviation.

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2}.$$

The reason we use $n-1$ instead of n is to further account for the imprecision in our estimates. statisticians say that 1 "degree of freedom" has already been lost, since we had to first estimate μ, before we have estimated σ.

Therefore, when the true mean μ and standard deviation are both unknown, we construct a $(100)(1-\alpha)\%$ CI using the formula

$$\left[\hat{\mu} \pm t_* \frac{\hat{\sigma}}{\sqrt{n}}\right]$$

instead of

$$\left[\hat{\mu} \pm z_* \frac{\sigma}{\sqrt{n}}\right]$$

as before. The two differences between these two formulas are

- We replaced σ with the estimate $\hat{\sigma}$ (since σ is unknown).
- We are using a t-score instead of a z-score.

Here, t_* is the value for which $\mathbb{P}(|T| \leq t_*) = 1-\alpha$, where T has Student's t-distribution with $n-1$ degrees of freedom. Here, we can use an online calculator (note that $v = n-1$ in the calculator).

We can then use these ideas to run statistical hypothesis tests.

(Example.) Suppose that we would like to determine if the usual body temperature of humans differs from 98.6 degrees F. Note that we do not know the true mean μ nor the true standard deviation σ.

Suppose we take a random sample of 130 temperatures (so we know all of the X_i's), and find that $\hat{\mu} = 98.25$ and $\hat{\sigma} = 0.73$. At the 0.05 (5%) "significance level," can we reject the **null hypothesis** that $\mu = 98.6$. This means that if we reject the hypothesis, then there's only a 5% chance we're making a mistake by doing so. In other words, if we reject the hypothesis that $\mu = 98.6$, then we can do so with the probability of 5% that we're making a mistake.

Under the null hypothesis, the statistic

$$T = \frac{\hat{\mu} - 98.6}{\sigma^2/\sqrt{n}}$$

is approximately a Student's t with $n-1 = 129$ degrees of freedom. Therefore, $\mathbb{P}(|T| \geq 1.978) = 5\%$. There's only a 5% chance that we will observe something as extreme at something like 1.978. So, that means that if we get a statistics that's larger than 1.978, then that means that there's only a 5% chance that it's really true that 98.6 is the true temperature.

Now, let's suppose that $\hat{\mu} = 98.25$ and $\hat{\sigma} = 0.73$ and so

$$T = \frac{98.25 - 98.6}{0.73/\sqrt{130}} \approx -5.47,$$

which is a lot more extreme (and so even more unlikely) than 1.978. Hence, under the null hypothesis, the chance of us observing this statistic $T \approx -5.47$ is (much) less than 5%. Therefore, at the $\alpha = 0.05$ significance level, we reject the null hypothesis, in favor of the **alternative hypothesis** that the true average temperature is lower than 98.6 degrees F.