LECTURE: CENTRAL LIMIT THEOREM AND POINT ESTIMATORS (I)

1. Central Limit Theorem

Recall:

If $Y_1, \cdots, Y_n$ are iid then the **sample mean** is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Usually the distribution of $\overline{Y}$ is unknown

The cool thing is that (a rescaled version of) $\overline{Y}$ is still *approximately* normal, provided n is large:

Central Limit Theorem:

Let $Y_1, \ldots, Y_n$ be iid with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$

$$\text{Let } Z_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

Then, as $n \to \infty$ the distribution of Z_n converges to the standard normal distribution

Note: Mathematically, this means that for all z

$$\lim_{n\to\infty} P(Z_n \leq z) = P(Z \leq z) \text{ where } Z \sim N(0,1)$$

Practically, this means that for a large sample size $\bar{Y}$ is roughly $N\left(\mu, \frac{\sigma^2}{n}\right)$
Usually $n \geq 30$ is enough

Example 1:

Suppose SAT scores usually have a mean of 60 and variance 64

A new version of the test is given to 100 students, and the mean score of those 100 students is 58.

How likely is it that there is something wrong with that new version?

We'll calculate $P(\bar{Y} \leq 58)$

By the Central Limit Theorem, since our sample size is large (≥ 30), $\overline{Y}$ is approximately normal with mean $\mu = 60$ and variance

$$\frac{\sigma^2}{n} = \frac{64}{100} \Rightarrow \frac{\sigma}{\sqrt{n}} = \frac{8}{10} = 0.8$$

Converting to the standard normal random variable:

$$\begin{aligned} P(\bar{Y} \leq 58) =& P\left(\frac{\bar{Y} - 60}{0.8} \leq \frac{58 - 60}{0.8}\right) \\ \approx& P\left(Z \leq \frac{58 - 60}{0.8}\right) \quad \text{(Central Limit Theorem)} \\ =& P(Z \leq -2.5) = F(-2.5) = 0.0062 \end{aligned}$$

This probability is so small that it is unlikely that it was drawn from a population with mean 60 and variance 64. Thus it is highly likely that there is something wrong with the test.

Example 2:

Service times for customers in a retail store are independent random variables with mean 1.5 minutes and variance 1.0 minutes.

Approximately what is the probability that 100 customers can be served in less than 2 hours?

Let Y_i be the service time for the ith customer, then

$$P\left(\sum_{i=1}^{100} Y_i \leq 120\right) = P\left(\frac{1}{100}\sum_{i=1}^{100} Y_i \leq \frac{120}{100}\right) = P(\bar{Y} \leq 1.20)$$

Since n is large, by the Central Limit Theorem, $\bar{Y}$ is approximately normally with mean $\mu = 1.5$ and variance $\frac{\sigma^2}{n} = \frac{1}{100} \Rightarrow \frac{\sigma}{\sqrt{n}} = 0.1$

Thus, converting to the standard normal random variable, we have:

$$\begin{aligned} P(\bar{Y} \leq 1.20) &\approx P\left(Z \leq \frac{1.20 - 1.5}{0.1}\right) \\ &= P\left(Z \leq \frac{-0.30}{0.1}\right) \\ &= P(Z \leq -3.0) = F(-3) = 0.0013 \end{aligned}$$

This probability is so small that it is virtually impossible to serve 100 customers in less than 2 hours.

Note: What makes this so nice is that we can use this even though the distributions of the service times is unknown. If we had to model this realistically, we might choose an exponential distribution, although we would have to change either the mean or the variance in that case.

2. Estimators

The whole point of statistics is to make inferences about a population based on data from a small sample of that population.

So far we have used $\overline{Y}$ to make a guess for the population mean μ and S^2 to make a guess about the variance σ^2

But there might be other parameters the distributions depend on. In the Chocolate/Vanilla ice cream scenario for example, a natural parameter is p, the proportion of the population who prefers Chocolate.

This is why it's useful to generalize the approach that we have so far.

Suppose we are studying a population whose distribution has a parameter which we will denote θ. This could be the population mean, population variance, proportion of people who prefer Chocolate etc.

We will take n iid samples $Y_1, \ldots, Y_n$ from our population.

Definition:

An **estimator** $\hat{\theta}$ is a function of $Y_1, \cdots, Y_n$ that gives us information about θ

(1) A **point estimator** produces a single number that is close to the parameter of interest

(2) An **interval estimate** produces an interval (often called a confidence interval) in which we believe our parameter of interest lies.

Note: We generally use hats for estimators. So the "hat" over the θ indicates that it is an estimator θ

3. Point Estimators

We have already met one point estimator, the sample mean $\bar{Y}$ which is an estimator for the population mean μ

Example 3:

Suppose you are polling n people out of a population and ask them if they prefer Chocolate over Vanilla.

Here the parameter of interest is p, the proportion of people who prefer chocolate

Let Y be the number of people in our sample who prefer Chocolate.

Then the sample proportion $\hat{p} = Y/n$ is an estimator for the population proportion p

What does it mean for an estimator to be good? Here we need some quantitative tools that measures "goodness"

Let $\hat{\theta}$ be an estimator for θ, notice $\hat{\theta}$ is a random variable.

In order for $\hat{\theta}$ to be an approximation for θ we should *at least* have $E(\hat{\theta}) = \theta$.

Definition:

$\hat{\theta}$ is **unbiased** if $E(\hat{\theta}) = \theta$ else $\hat{\theta}$ is **biased**

Think about it! What if $\theta = 10$ but on average, your estimator $\hat{\theta}$ gives you 30? There would be something iffy with that estimate!

Definition:

The **bias** of $\hat{\theta}$ is given by $\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$

Let's look at the estimators we have seen so far:

Example 4:

(a) Is the sample mean $\bar{Y}$ biased?

The sample mean $\hat{\theta} = \bar{Y}$ is an estimator for the population mean $\theta = \mu$ We have previously shown that $E(\bar{Y}) = \mu$, thus the sample mean is an unbiased estimator for the population mean.

(b) Is the sample proportion biased?

What about the sample proportion? Suppose we poll n people, and let Y be the number of people in our sample prefer Chocolate. Here $Y \sim$ Binom (n, p)

For our estimator $\hat{p} = Y/n$,

$$E(\hat{p}) = E\left(\frac{Y}{n}\right) = \frac{1}{n}E(Y) = \frac{np}{n} = p$$

Since $E(\hat{p}) = p$, this estimator is unbiased as well.