Sampling and Estimation

ECON 340: Economic Research Methods — Instructor: Div Bhagia

We often want to make inferences about a population. But a lot of times, it is only possible to collect data from some individuals in the population. Therefore, we collect data from a smaller subset of the population, called a sample. However, the sample must be selected carefully to ensure that the inferences we make from the sample accurately reflect the characteristics of the population.

One way to achieve a representative sample is through the use of *random sampling*. In a random sample, each individual in the population has an equal chance of being selected for the sample. This helps reduce bias that could be introduced if individuals were chosen based on specific characteristics.

Sample statistics, such as the sample mean or sample standard deviation, are quantities calculated from a sample of data. Because the sample is only a subset of the entire population, sample statistics vary from sample to sample. This variability means that *sample statistics are random variables*.

To see why sample statistics are random variables, consider the sample mean as an example. Suppose we take multiple random samples of the same size from a population and calculate the mean for each sample. We expect these sample means to vary from sample to sample due to the variability inherent in the sampling process. Thus, the sample mean is a random variable since it can take on different values depending on the sample that is selected.

1 Distribution of Sample Mean

Let $X_1, X_2, ..., X_n$ denote independent random draws (random sample) from a population with mean μ and variance σ^2. We can calculate the sample mean as follows:

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

We can show that the expectation and the variance of the sample mean are given by[1]:

$$E(\bar{X}) = \mu \qquad Var(\bar{X}) = \frac{\sigma^2}{n}$$

The expected value of the sample mean is equal to the population mean, which means that if we were to take an infinite number of random samples and calculate the mean of each sample, the average of these sample means would be equal to the population mean.

The variance of the sample mean refers to the amount of variability or spread that we would expect to see in the sample means that we would obtain if we were to take multiple random samples from a population.

The variance of the sample mean is equal to the population variance divided by the sample size. From this formula, we can see that when sample size increases, the variance of the sample mean decreases, and the sample mean becomes a more precise estimator of the population mean. This is because by increasing the sample size, we can reduce the effects of random variation and sampling error.

On the other hand, variance of the sample mean increases with the variance of the population. This means that when we take repeated random samples of a fixed size from a highly variable population, we are likely to observe more variability in the sample means from sample to sample.

The distribution of the sample mean is normal if *either* of the following is true:

- The underlying population is normal
- The sample size is large, say $n \geq 100$

The first one follows from the sample mean being a linear combination of normally distributed variables. The latter is implied by the *Central Limit Theorem*.

[1] Derivations at the end.

Central Limit Theorem

Central Limit Theorem (CLT) states that if $X_1, X_2, .., X_n$ are drawn randomly from a population with mean μ and variance σ^2, sample mean $\bar{X}$ is normally distributed with mean μ and variance σ^2/n as long as n is large.

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

In other words, CLT states that as we increase the sample size, the distribution of the sample means tends to approximate a normal distribution, regardless of the underlying distribution of the population.

2 Estimators

An estimator $\hat{\theta}$ for the population parameter θ is said to be *unbiased* if

$$E(\hat{\theta}) = \theta$$

Examples: $\bar{X}$ for μ, s^2 for σ^2 But lots of estimators are unbiased. For example, say our estimator is X_1, then $E(X_1) = \mu$. This doesn't sound right. What else should we be looking for?

When choosing between two unbiased estimators $\hat{\theta_1}$ and $\hat{\theta_2}$, we prefer the lower variance estimator. We say the lower variance estimator is more *efficient*.

3 Confidence Intervals

Since the sample mean is a random variable, we cannot expect it to be exactly equal to the true population mean in any particular sample. However, if we have a large and random sample, we can use the sample mean as an estimate of the population mean with some degree of confidence. The sample variance measures how much the sample mean deviates from the population mean on average. Confidence intervals are another way to summarize this uncertainty by providing a range of values that contains the population mean with a certain probability.

Essentially our goal is to create an interval around the sample mean that gives us a range of plausible values for the population mean. We can have confidence intervals of varying levels of confidence, most common are 90%, 95%, or 99%. The level of confidence is the probability that a calculated confidence interval contains the true population parameter.

Say we are interested in creating a 95% confidence interval for the true population mean. If we can conclude that $\bar{X} \sim N(\mu, \sigma^2_{\bar{X}})$, then

$$Z = \frac{\bar{X} - \mu}{\sigma^2_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma^2/\sqrt{n}} \sim N(0, 1)$$

Note that from the formula for the variance of the sample mean $\sigma^2_{\bar{X}} = \sigma^2/\sqrt{n}$.

From the standard normal table we can see that $Pr(-1.96 < Z < 1.96) = 0.95$. This implies that

$$P\left(\mu - 1.96 \cdot \frac{\sigma^2}{\sqrt{n}} < \bar{X} < \mu + 1.96 \cdot \frac{\sigma^2}{\sqrt{n}}\right) = 0.95$$

The above implies that the sample mean $\bar{X}$ is within 1.96 standard deviations of the population mean μ with a 95% probability. We can now push this reasoning further and say that in that case, it must be that the population mean μ is within 1.96 standard deviations of any realization of the sample mean $\bar{X}$ with a 95% probability. In other words, if we observe a sample mean $\bar{x}$, then there is a 95% chance that the population mean μ is within 1.96 standard deviations of $\bar{X}$. This gives us a way to construct a 95% confidence interval for μ based on $\bar{x}$ as follows:

$$\bar{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Confidence Interval: Known Population Variance

Let $z_{\alpha/2}$ be the z-value that leaves area $\alpha/2$ in the upper tail of the normal distribution. Then $1 - \alpha$ confidence interval is given by

$$\bar{x} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{Margin of Error}}$$

The margin of error is influenced by two main factors: the standard deviation of the sample mean, which in turn is affected by both the population standard deviation and the sample size, and the level of confidence. As the level of confidence increases, the margin of error also increases, resulting in a wider interval. If greater confidence is desired, a wider range of values must be considered. Additionally, if the sample variance is higher, there is more uncertainty, which leads to a wider interval.

Unknown Population Variance

Until now, we have made the assumption that the population variance is known, but this is often not the case in practice. However, we can rely on the sample variance as an unbiased estimator for the population variance. As a result, we can no longer construct the Z statistic, but we can use the T statistic instead, which is essentially the same as the Z statistic, but utilizes the sample standard deviation instead of the population standard deviation. This T statistic is defined as:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

It has been shown that this statistic follows a t distribution with $n - 1$ degrees of freedom. The t-distribution is similar to the normal distribution, but it has thicker tails to account for the greater uncertainty in smaller sample sizes. In larger sample sizes, the t-distribution can be approximated by the standard-normal distribution.

To construct a confidence interval using the T statistic, we can use the same approach as before, but we now need to use the critical value for the t-distribution, denoted as $t_{\alpha/2,n-1}$.

Confidence Interval: Unknown Population Variance

Let $t_{\alpha/2,n-1}$ be the t-value that leaves area $\alpha/2$ in the upper tail of the t-distribution with $n - 1$ degrees of freedom. Then $1 - \alpha$ confidence interval is given by

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

However, if the sample size is larger than or equal to 100, the t-distribution can be

approximated by the standard-normal distribution. In this case, we can continue to use the standard-normal table for the critical values.

4 Hypothesis Testing

So far, we have discussed that when we use a sample to make inferences about the population, the sample statistic is not necessarily equal to the population parameter. However, we might be interested in knowing how plausible it is that the population parameter takes a particular value given our realization of the sample mean. To do this, we can formally test the hypothesis that the population parameter takes a particular value.

Say, we are interested in determining whether the population mean μ is equal to a particular value μ_0. We found a sample mean of $\bar{x}$. We can formally test our hypothesis $\mu = \mu_0$ by following a set of steps. However, before we proceed, we need to choose a significance level. The significance level is the probability of rejecting the null hypothesis when it is actually true. Common significance levels are 0.01, 0.05, or 0.1, which correspond to a 1%, 5%, or 10% chance of rejecting the null hypothesis when it is actually true.

The steps for testing the hypothesis $\mu = \mu_0$ are as follows:

1. *Formulate the null and alternative hypothesis*

$$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$$

   The null hypothesis assumes that the population mean is equal to the specified value μ_0, while the alternative hypothesis assumes that it is not.

2. *Determine the distribution of the test statistic under the null.* If we can conclude that the sample mean is normally distributed around the true population mean, then *under the null* $\bar{X} \sim N(\mu_0, \sigma^2/n)$. This implies that under the null, and T-statistic is distributed according to t_{n-1} (or the Z statistic is distributed according to the

standard normal distribution).

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

3. *Determine the rejection region.* We reject the null hypothesis when the calculated test statistic falls in the $\alpha\%$ most extreme outcomes of its distribution. This is because we started by assuming the null hypothesis to be true, and if we find a value that is too far in the tails, it suggests that the null hypothesis is unlikely to be true. Thus, we reject the null hypothesis in favor of the alternative hypothesis.

4. Calculate the test statistic (T or Z depending on if you know the population variance or not) and reject or fail to reject the null.

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}}$$

   If the test statistic falls in the rejection region, we reject the null hypothesis in favor of the alternative hypothesis. So reject the null if $|t| > t_{n-1,\alpha/2}$ where $t_{n-1,\alpha/2}$ is the critical value of the t-distribution with $n-1$ degrees of freedom at a significance level of α.

p-value

In hypothesis testing, the p-value is the probability of observing a test statistic as extreme or more extreme than the one calculated from the sample data, assuming the null hypothesis is true. We can calculate it by finding the area in the tails of the distribution beyond the absolute value of the observed test statistic and doubling it. In particular,

$$p = 2Pr(T \geq |t||H_0 : \mu = \mu_0)$$

To put it simply, the p-value measures the strength of evidence against the null hypothesis. A small p-value indicates that the observed data is unlikely to have occurred by chance alone, and provides evidence in favor of the alternative hypothesis. On the

other hand, a large p-value suggests that the observed data could have occurred by chance, and there is not enough evidence to reject the null hypothesis.

We can use the significance level α and p-value to make a decision about whether to reject or fail to reject the null hypothesis. In particular, if the p-value is smaller than the significance level, we reject the null hypothesis, and if the p-value is greater than or equal to the significance level, we fail to reject the null hypothesis.

It is important to note that the p-value does not give any information about the size of the effect or the practical significance of the result. It only provides information on the statistical significance of the result.

Additional Derivations

The expectation of the sample mean can be derived as follows:

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}n\mu = \mu \end{aligned}$$

The variance of the sample mean can be derived as follows:

$$\begin{aligned} Var(\bar{X}) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$