Econ 103: Introduction to Econometrics
Week 1

Filipe Fiedler

Fall 2025

1 Parameters and estimators

In Economics, we are often interested in learning about some characteristic of a population. For example, we might want to know the average income of all households in the US. However, it is usually impractical or impossible to collect data from the entire population. Instead, we often collect a sample of data from the population and use it to estimate the characteristic of interest.

We can define the income of all households in the US as a random variable X. The average income of all households in the US is then the expected value of X, denoted by $\mu = \mathbb{E}[X]$. Since we do not know the distribution of X, we do not know the value of μ. However, we can collect a sample of data from the population, say $x_1, x_2, \ldots, x_n$, where $x_i \in \mathbb{R}$ is the income of household i. We can use this realized sample to estimate μ.

Before we observe the data, we can think of $X_1, X_2, \ldots, X_n$ as random variables, while $x_1, x_2, \ldots, x_n$ as their realizations. I will refer to $X_1, X_2, \ldots, X_n$ as the sample while $x_1, x_2, \ldots, x_n$ is the realized sample or sample of data. This distinction allows us to define estimators and to study their properties.

Parameter: Numerical characteristic of a distribution of one or more random variables. Examples: expected value μ, variance σ^2 and lower quartile p_{25}. Example with joint distributions: $\text{Corr}(X, Y)$ or $\text{Cov}(X, Y)$

Estimator: Random variable that is a function of the sample $(X_1, \ldots, X_n)$ and, when applied to a realized sample, estimates an unknown parameter. Examples: $\bar{X}_n$ for the expected value μ and S_n^2 for the variance σ^2.

Estimate: Realization of an estimator given a realization of sample of data. It is a real number function of the data $x_1, x_2, \ldots, x_n$. Examples: $\bar{x}_n$ or s_n^2.

We often want to know if an estimator is a good estimator for a parameter. There are two common properties that we can use to evaluate the quality of an estimator: unbiasedness and consistency.

Unbiased Estimator: An estimator $\tilde{\theta}$ is unbiased for the parameter θ if $\mathbb{E}[\tilde{\theta}] = \theta$. Example: $\mathbb{E}[\bar{X}_n] = \mu$

Intuitevely, imagine that we have B realized samples of size n from the population, where B is a large number. We can compute the estimate $\hat{\theta}_b$ for each realized sample $b = 1, 2, \ldots, B$. If we take the average of all estimates, we get $\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b$. If $\hat{\theta}$ is an unbiased estimator, then this average will tend to be very close to the true parameter θ.

Consistent Estimator: An estimator $\tilde{\theta}$ is consistent for the parameter θ if the distribution of $\tilde{\theta}$ converges to θ when the sample size goes to infinity. Example: Since $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \xrightarrow{n\to\infty} 0$, then distribution of $\bar{X}_n$ converges to μ

Intuitevely, as we collect more and more data, the estimate $\hat{\theta}$ will tend to be closer and closer to the true parameter θ.

1.1 Common estimators

Estimator for the mean:

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

This is the most common estimator for the mean. It is unbiased and consistent.

Estimator for the variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

This is the most common estimator for the variance. It is unbiased and consistent. Dividing by $n-1$ instead of n makes the estimator unbiased. Sometimes, we use the estimator $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$, which is biased but consistent.

Estimator for the covariance:

$$\widehat{\text{Cov}}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)\left(Y_i - \bar{Y}_n\right)$$

New estimators: A general method for constructing estimators in iid samples is to use the sample correspondent of the definition of the parameter. For example, if $\theta = \mathbb{E}[g(X)]$ for some continuous function g, then we can use the estimator $\tilde{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. One can show that this estimator is consistent. This is the basis for the (biased) estimators of the variance and the covariance.

1.2 Other important concepts

Standard Error: The standard error of an estimator $\tilde{\theta}$ is the standard deviation of its distribution, denoted by $\text{SE}(\tilde{\theta}) = \sqrt{\text{Var}(\tilde{\theta})}$. The standard error measures the variability of the estimator across different samples. A smaller standard error indicates that the estimator is more precise.

Independence: Two random variables X and Y are independent if the occurrence of one does not affect the probability distribution of the other. In other words, knowing the value of X does not provide any information about the value of Y, and vice versa. Formally, X and Y are independent if for all x and y, $P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y)$.

Random Sample: A random sample is a set of observations drawn from a population such that each observation is independent and identically distributed (i.i.d.) according to the same probability distribution as the population.

2 Central Limit Theorem

Suppose we have a random sample $X_1, \ldots, X_n$. The Central Limit Theorem (CLT) states that, for a large enough sample size n, the distribution of the sample mean $\bar{X}_n$ will be approximately normal, regardless of the distribution of the original random variable X. More formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and finite variance σ^2, then:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

as $n \to \infty$.

This means that, for large n, we can use the normal distribution to approximate the distribution of the sample mean. This is very useful because we can use the properties of the normal distribution to know how much uncertainty there is in our estimator and test if we have statistical evidence to reject some assertions about the data.

2.1 How large is large enough?

Not in the exam There is no hard and fast rule for how large n needs to be for the approximation provided by the CLT to be accurate. If the distribution of X has finite support, a small sample size is enough to give good approximations, like $n = 50$. If X has infinite support and its distribution is heavily skewed or has heavy tails, a larger sample size may be needed for the CLT to provide a good approximation. We can estimate the skewness and the kurtosis to check how far the distribution is from a normal distribution. Sometimes, we might want to use a transformation of the data to make it more normal-like before applying the CLT.

2.2 Beyond the mean

Not in the exam The CLT can be extended to other statistics beyond the sample mean by using properties of independence, properties of the normal distribution, continuous mapping theorem and the Delta Method. For example, we can use the CLT to derive the asymptotic distribution of the sample variance S_n^2 by defining $Y_i = (X_i - \mu)^2$ and applying the CLT to the sample mean of Y_i.

We can also use the CLT to derive the asymptotic distribution of the sample correlation $\widehat{\text{Corr}}(X, Y)$ by applying the continuous mapping theorem and the Delta Method to the joint distribution of $\bar{X}_n$, $\bar{Y}_n$, S_X^2 and S_Y^2.

3 Hypothesis Tests

A hypothesis test is a statistical method used to make affirmatives about the value of a parameter based on a sample of data. Imagine that we want to know if the average income of all households in the US is equal to \$100,000. We can collect a sample of data from the population and use it to test this hypothesis. Our estimate might be different from \$100,000 just by random chance, so we need a way to determine if the difference is statistically significant.

The basic idea is to formulate two competing hypotheses: the null hypothesis ($H_0 : \mu = \$100,000$) and the alternative hypothesis ($H_1 \neq \$100,000$) and a level of significance α. The level of significance is the risk the researcher is willing to take to reject the null hypothesis when it is actually true. In Economics, we often use $\alpha = 0.05$, but when an enginner is testing if the probability of a bridge falling is greater than a given threshold, they might use $\alpha = 0.001$.

Using the CLT, we can find a statistic for which the distribution is known under the null hypothesis and then compute the value of this statistic using our sample of data. If the value of the statistic is very unlikely under the null hypothesis, we have evidence against the null hypothesis and in favor of the alternative hypothesis. The threshold for what is considered "very unlikely" is determined by the level of significance α.

3.1 Example: Testing the mean

We want to test $H_0 : \mu = \mu_0 = \$100,000$ vs $H_1 : \mu \neq \$100,000$ at level of significance $\alpha = 0.05$. We collect a sample of size $n = 100$ and compute the sample mean $\bar{x}_n = \$105,000$ and the sample standard deviation $s_n = \$10,000$. The theory gives us that, if H_0 is true, then:

$$T = \frac{\bar{X}_n - \mu_0}{S_n/\sqrt{n}} \approx N(0,1)$$

Thus, we can compute the test statistic:

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}} = \frac{105,000 - 100,000}{10,000/\sqrt{100}} = \frac{15,000}{3,000} = 5$$

Values close to 5 are very unlikely in a $N(0,1)$ distribution. That gives us evidence against the null hypothesis. More formally, we can compute a critical value c such that $P(|Z| > c) = \alpha$, where $Z \sim N(0,1)$. For $\alpha = 0.05$, we have $c \approx 1.96$. Since $|t| = 5 > 1.96$, we reject the null hypothesis. In words, we have statistical evidence that the average income of all households in the US is different from \$100,000.

We can perform this test using the p-value approach as well. The p-value is the probability of observing a test statistic as extreme or more extreme than the one observed, assuming that the null hypothesis is true. In our example, we can compute the p-value as follows:

$$\text{p-value} = P(|Z| > |t|) = P(|Z| > 5)$$

where $Z \sim N(0,1)$. Using a standard normal table or a statistical software, we find that the p-value is very small (less than 0.0001). Since the p-value is less than the level of significance $\alpha = 0.05$, we reject the null hypothesis in favor of the alternative hypothesis.

4 Confidence Intervals

A confidence interval is a range of values that is likely to contain the true value of a parameter with a certain level of confidence. We can denote a confidence interval as (L, U), where L is the lower bound and U is the upper bound. The level of confidence is denoted by $1 - \alpha$, where α is the level of significance used in hypothesis testing.

L and U are random variables that depend on the sample, so the confidence interval is also a random variable. The confidence interval is constructed such that, if we were to repeat the sampling process many times, a proportion $1-\alpha$ of the confidence intervals would contain the true parameter value.

$$P\left(L \leq \theta \leq U\right) \approx 1 - \alpha$$

where θ is the true parameter value. The approximation becomes exact as the sample size n goes to infinity.

We can construct a confidence interval for the mean using the CLT. If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then a $(1 - \alpha)$ confidence interval for μ is given by:

$$\left(\bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution.

4.1 Example: Confidence interval for the mean

We want to construct a 95% confidence interval for the mean income of all households in the US. We collect a sample of size $n = 100$ and compute the sample mean $\bar{x}_n = \$105,000$ and the sample standard deviation $s_n = \$10,000$. Using the formula for the confidence interval, we have:

$$\begin{aligned}
[l, u] &= \left[105,000 - 1.96\frac{10,000}{\sqrt{100}}, 105,000 + 1.96\frac{10,000}{\sqrt{100}}\right] && (1)\\
&= [105,000 - 1.96 \times 1,000, 105,000 + 1.96 \times 1,000] && (2)\\
&= [105,000 - 1,960, 105,000 + 1,960] && (3)\\
&= [103,040, 106,960] && (4)
\end{aligned}$$

We can interpret this confidence interval as follows: we are 95% confident that the true mean income of all households in the US is between \$103,040 and \$106,960. This means

that if we were to repeat the sampling process many times and construct a 95% confidence interval for each sample, approximately 95% of those intervals would contain the true mean income.

To be clear, this does not mean that there is a 95% probability that the true mean income is between \$103,040 and \$106,960. That probability is either one or zero, since the true mean income is a fixed value. Instead, it means that the **method** we used to construct the confidence interval has a 95% success rate in capturing the true mean income. The interval $[103,040, 106,960]$ can be one of the 5% of intervals that do not contain the true mean income.

5 Monte Carlo Simulations

We often do not know the distribution of an estimator, since the distribution of the random variable of interest, X, is unknown. However, to understand the properties of an estimator $\tilde{\theta}$, we can use Monte Carlo simulations. The idea is to simulate B samples from a known distribution and compute the estimate $\hat{\theta}$ for each sample. We can then study the distribution of the estimator across all samples.

5.1 Example: Monte Carlo simulation for the mean

Imagine we have a random variable X that follows an Exponential distribution with parameter $\lambda = 1$. The expected value of X is $\mu = 1/\lambda = 1$. We want to estimate μ using the sample mean $\bar{X}_n$. We can use a Monte Carlo simulation to calculate the distribution of $\bar{X}_n$ and verify it is an unbiased estimator for μ and that its distribution is approximately normal for large n. If we increase n, we can also verify that the distribution of $\bar{X}_n$ becomes more concentrated around μ, which is the intuition behind the result that $\bar{X}_n$ is a consistent estimator for μ.

Pseudocode:

```
Set B = 1000
Set n = 100
Set lambda = 1

For b in 1 to B:
    Simulate a sample of size n from an Exponential dist. with parameter lambda
    Compute the sample mean x_bar_b
    Store x_bar_b in a vector x_bar
```

```
Compute the average of all x_bar
Plot the histogram of all x_bar
```

The unbiasedness of the sample mean can be verified by checking if the average of all x_bar is close to the true mean of the Exponential distribution, which is $1/\lambda = 1$. The consistency of the sample mean can be verified by increasing n and checking if the variance of all x_bar decreases.

6 Empirical Questions

These exercises were designed to help you understand the concepts covered in this week's notes and to start our study of R.

6.1 Question 1

Empirical verification of the CLT.

Consider the random variable X that follows an Exponential distribution with parameter $\lambda = 1$. Run a Monte Carlo simulation with $B = 1000$ and $n = 2$ and calculate, for each sample of data, the statistic $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$. Plot the histogram of this statistic and overlay the density of a $N(0, 1)$ distribution. Repeat the exercise with $n = 10$ and $n = 50$. What do you observe?

6.2 Question 2

Empirical verification of unbiasness and consistency of the sample variance.

Consider the random variable X that follows a Chi-squared distribution with 50 degrees of freedom. The variance of this distribution is 100. Fix n. Run a Monte Carlo simulation with $B = 1000$ and $N = n$ and calculate, for each sample of data, the sample variance. Calculate the difference between the average of all sample variances and the true variance of X. Plot the histogram of the sample variances. Repeat the exercise with $n = 10, 100, 1000$. What do you observe?

6.3 Question 3

Confidence Intervals for the mean.

Consider the random variable X that follows a Normal distribution with mean $\mu = 5$ and variance $\sigma^2 = 4$. Run a Monte Carlo simulation with $B = 1000$ and $n = 100$. For each sample of data, compute the 90% confidence interval for the mean usind the sample variance. Calculate the proportion of confidence intervals that contain the true mean $\mu = 5$.

6.4 Question 4

Speed of convergence to a normal distribution.

Consider the random variables X, Y. X follows an Uniform distribution with support $[50, 100]$. Y follows a Pareto distribution with parameter $\alpha = 2.01$, such that $\mu = \frac{\alpha}{\alpha - 1}$. Plot the density of both distributions.

Run a Monte Carlo simulation with $B = 1000$ and $n = 10, 100, 1000$. For each sample of data, compute the statistic $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ and $\frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$. Plot the histogram of both statistics and overlay the density of a $N(0, 1)$ distribution. What do you observe?

7 Theoretical Questions

7.1 Question 1

Suppose a warehouse has 64 machines that each take some time to complete a routine safety check. Let Y_i be the time (in minutes) it takes machine i to finish the check, and assume all Y_i are independent and identically distributed. Also assume that the mean of each Y_i is 45 minutes and the standard deviation is 12 minutes.

(a) Let $\bar{Y}$ be the average time it takes for the machines to finish their checks. What is the probability that, on average, the checks take more than 47 minutes to complete? *(Hint: Use the approximation* $\bar{Y} \sim N(E[Y], \text{Var}(Y)/n)$ *to obtain an approximate probability.)*

(b) What is the probability that, on average, the checks take between 44 and 46 minutes?

(c) Look back at your answers to (a) and (b). Are these exact probabilities? Could you compute exact probabilities with the information given? Justify your answer.

8 Solutions to Exercises

8.1 Question 1

Suppose a warehouse has 64 machines that each take some time to complete a routine safety check. Let Y_i be the time (in minutes) it takes machine i to finish the check, and assume all Y_i are independent and identically distributed. Also assume that the mean of each Y_i is 45 minutes and the standard deviation is 12 minutes.

(a) Let $\bar{Y}$ be the average time it takes for the machines to finish their checks. What is the probability that, on average, the checks take more than 47 minutes to complete? *(Hint: Use the approximation* $\bar{Y} \sim N(E[Y], \text{Var}(Y)/n)$ *to obtain an approximate probability.)*

Solution: We have that $E[Y] = 45$ and $\text{Var}(Y) = 12^2 = 144$. Since we have $n = 64$ machines, we can use the CLT to approximate the distribution of $\bar{Y}$ as follows:

$$\bar{Y} \sim N\left(45, \frac{144}{64}\right) = N(45, 2.25)$$

We want to find $P(\bar{Y} > 47)$. We can standardize this probability using the properties of the normal distribution:

$$\begin{aligned} P(\bar{Y} > 47) &= P\left(\frac{\bar{Y} - 45}{\sqrt{2.25}} > \frac{47 - 45}{\sqrt{2.25}}\right) \\ &= P\left(Z > \frac{2}{1.5}\right) \\ &= P(Z > 1.33) \end{aligned}$$

where $Z \sim N(0, 1)$. Using a standard normal table or a statistical software, we find that $P(Z > 1.33) \approx 0.0918$. Therefore, the probability that, on average, the checks take more than 47 minutes to complete is approximately 0.0918.

(b) What is the probability that, on average, the checks take between 44 and 46 minutes?

Solution: We want to find $P(44 < \bar{Y} < 46)$. We can standardize this probability

using the properties of the normal distribution:

$$\begin{aligned} P(44 < \bar{Y} < 46) &= P\left(\frac{44-45}{\sqrt{2.25}} < \frac{\bar{Y}-45}{\sqrt{2.25}} < \frac{46-45}{\sqrt{2.25}}\right) \\ &= P\left(-\frac{1}{1.5} < Z < \frac{1}{1.5}\right) \\ &= P(-0.67 < Z < 0.67) \end{aligned}$$

where $Z \sim N(0,1)$. Using a standard normal table or a statistical software, we find that $P(-0.67 < Z < 0.67) \approx 0.496$. Therefore, the probability that, on average, the checks take between 44 and 46 minutes is approximately 0.496.

(c) Look back at your answers to (a) and (b). Are these exact probabilities? Could you compute exact probabilities with the information given? Justify your answer.

Solution: These are not exact probabilities. We used the Central Limit Theorem to approximate the distribution of $\bar{Y}$ as a normal distribution. The CLT provides an approximation that becomes more accurate as the sample size n increases. We can not compute the exact probabilities with the information given, since we do not know the exact distribution of the individual Y_i. If we knew the exact distribution of Y_i, we could compute the exact distribution of $\bar{Y}$ and thus the exact probabilities.