Handout 3: Random Sampling & Large Sample Approximations

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Population vs. Sample

Question 1.1: What is the difference between a *population parameter* and a *sample statistic*? Give an example of each using this earnings data.

A **population parameter** is a fixed, true characteristic of the entire population. It is typically **unknown** because we cannot observe the whole population.

A **sample statistic** is a value computed from sample data to estimate a population parameter. Because it depends on which observations happen to be drawn, it is a **random variable**.

Examples:

- Population parameter: $\mu_Y = E[Y]$, the true mean weekly earnings of all 100,000 workers
- Sample statistic: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the sample mean computed from n randomly drawn workers

Question 1.2: Plot a histogram of the population earnings. Describe the shape of the distribution.

The histogram shows a **right-skewed** (positively skewed) distribution. The distribution is **not symmetric**: it has a long right tail, meaning there are some workers with very high earnings pulling the distribution to the right. Most workers earn amounts clustered toward the lower end, with the mean pulled above the median by the right tail.

2 Random Sampling

Question 2.1: What does it mean for a sample to be **i.i.d.**? Why is random sampling important?

Random variables $Y_1, Y_2, \ldots, Y_n$ are **independently and identically distributed (i.i.d.)** if:

1. **Identically distributed**: Each Y_i comes from the same probability distribution (same population)
2. **Independent**: The value of any Y_i provides no information about any other Y_j $(i \neq j)$

Random sampling ensures the i.i.d. property because:

- Each observation is equally likely to be drawn from the population ⇒ identically distributed
- Each draw does not depend on previous draws ⇒ independent

The i.i.d. assumption is crucial because it allows us to derive properties of estimators (like the sample mean) and to apply the Law of Large Numbers and the Central Limit Theorem.

Question 2.2: Draw random samples and compute sample means.

(a) Neither sample mean is likely to be exactly equal to the population mean. This is because random sampling introduces **sampling variability**: each sample is just one possible realization from the population, so $\bar{Y}$ will differ from μ_Y by chance.

(b) The sample mean with $n = 100$ is likely to be closer to the population mean than the sample mean with $n = 10$. This is because $\text{Var}(\bar{Y}) = \sigma_Y^2/n$: larger samples have smaller variance, meaning the sample mean is more tightly concentrated around μ_Y.

3 The Sample Mean as a Random Variable

Question 3.1: Why is the sample mean $\bar{Y}$ a **random variable**?

The sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a random variable because it is a function of random variables $Y_1, \ldots, Y_n$. Every time we draw a new random sample from the population, we get different values of $Y_1, \ldots, Y_n$, and therefore a different value of $\bar{Y}$.

Because $\bar{Y}$ is a random variable, it has its own **probability distribution** (called the *sampling distribution*), its own expected value $E[\bar{Y}]$, and its own variance $\text{Var}(\bar{Y})$.

Question 3.2: Show mathematically that $E[\bar{Y}] = \mu_Y$ and $\text{Var}(\bar{Y}) = \sigma_Y^2/n$.

(a) Expected value:

$$\begin{aligned} E[\bar{Y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] &= \frac{1}{n}E\left[\sum_{i=1}^{n} Y_i\right] \\ &= \frac{1}{n}\left(E[Y_1] + E[Y_2] + \cdots + E[Y_n]\right) \\ &= \frac{1}{n}\left(\mu_Y + \mu_Y + \cdots + \mu_Y\right) = \frac{1}{n}(n \cdot \mu_Y) = \boxed{\mu_Y} \end{aligned}$$

This says that $\bar{Y}$ is an **unbiased** estimator of μ_Y: on average, across all possible samples, $\bar{Y}$ equals μ_Y.

(b) Variance:
Because Y_i and Y_j are **independent** for $i \neq j$:

$$\begin{aligned} \text{Var}(\bar{Y}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) &= \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n^2}\left(\text{Var}(Y_1) + \text{Var}(Y_2) + \cdots + \text{Var}(Y_n)\right) \\ &= \frac{1}{n^2}(n \cdot \sigma_Y^2) = \boxed{\frac{\sigma_Y^2}{n}} \end{aligned}$$

The key step uses the fact that independence implies $\text{Var}(Y_i + Y_j) = \text{Var}(Y_i) + \text{Var}(Y_j)$ (no covariance terms).

Question 3.3: What happens to the standard error as n increases?

The **standard error** of $\bar{Y}$ is:

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

As n increases, the standard error **decreases**:

- $n = 10$: SE $= \sigma_Y/\sqrt{10} \approx \sigma_Y \times 0.316$

- $n = 100$: SE $= \sigma_Y/\sqrt{100} = \sigma_Y \times 0.10$
- $n = 1000$: SE $= \sigma_Y/\sqrt{1000} \approx \sigma_Y \times 0.032$

This makes sense: larger samples contain more information about the population, so our estimate of the mean becomes more precise. Note that the standard error decreases at the rate $1/\sqrt{n}$—to cut the standard error in half, you need to **quadruple** the sample size.

4 Law of Large Numbers

Question 4.1: State the Law of Large Numbers in your own words.

Law of Large Numbers (LLN): If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2 < \infty$, then:

$$\bar{Y} \xrightarrow{p} \mu_Y$$

In plain language: as the sample size n grows, the sample mean $\bar{Y}$ gets arbitrarily close to the true population mean μ_Y. Larger samples give more accurate estimates.

Conditions required:

1. The observations must be **i.i.d.** (independently and identically distributed)
2. The variance σ_Y^2 must be **finite**

Question 4.2: Describe the pattern as sample size increases.

(a) The deviations $(\bar{Y} - \mu_Y)$ shrink toward zero as n increases. With small samples (e.g., $n = 10$ or $n = 25$), the sample mean can be quite far from the population mean. With large samples (e.g., $n = 10{,}000$ or $n = 50{,}000$), the sample mean is very close to μ_Y.

(b) The exact sample size depends on the random draw, but typically by $n = 500$ to $n = 1{,}000$, the deviation is quite small (within a few dollars of the population mean). This illustrates the LLN: larger samples yield sample means that are closer to μ_Y.

Question 4.3: Explain the LLN plot.

The plot illustrates two key results:

Law of Large Numbers: The dots (sample means) converge toward the red dashed line (population mean) as n increases along the horizontal axis. For small n, the dots are scattered far from μ_Y; for large n, they cluster tightly around it.

Variance formula: The vertical spread of the dots decreases as n increases. This directly illustrates $\text{Var}(\bar{Y}) = \sigma_Y^2/n$—the variance of $\bar{Y}$ shrinks proportionally to $1/n$. At each sample size, drawing 10 different samples shows how much variability remains; this variability clearly diminishes with larger n.

5 Central Limit Theorem

Question 5.1: State the Central Limit Theorem. Why is it "remarkable"?

Central Limit Theorem (CLT): If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2 < \infty$, then:

$$\bar{Y} \overset{a}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

or equivalently:

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \xrightarrow{d} N(0,1)$$

It is "remarkable" because it applies **regardless of the original distribution of** Y. The population can be skewed, bimodal, discrete, or any other shape—as long as it has a finite mean and variance, the sampling distribution of $\bar{Y}$ will be approximately normal for large n. This is why the normal distribution plays such a central role in statistics.

Question 5.2: Describe the sampling distributions for different n.

(a) **Shape of sampling distribution:**

- $n = 5$: Still visibly **right-skewed**, reflecting the skewness of the population. The CLT approximation is poor for such a small sample.
- $n = 30$: Approximately **symmetric and bell-shaped**. Some slight skewness may remain, but the distribution is close to normal.
- $n = 200$: Very close to a **perfect normal distribution**. The CLT approximation is excellent.

(b) Even though the population is right-skewed, the sampling distribution of $\bar{Y}$ becomes symmetric because $\bar{Y}$ is an **average of n random draws**. By averaging, extreme values are diluted. The CLT guarantees that this averaging process produces an approximately normal distribution for large n.

(c) The spread (variance) **decreases** as n increases. This is because $\text{Var}(\bar{Y}) = \sigma_Y^2/n$, so larger n means smaller variance. The histograms become narrower and more concentrated around μ_Y.

Question 5.3: How well does the normal curve fit?

At $n = 100$, the normal curve fits the histogram very well. The theoretical $N(\mu_Y, \sigma_Y^2/100)$ density closely matches the empirical sampling distribution, confirming that the CLT provides an excellent approximation at this sample size. This means we can use the normal distribution to calculate probabilities about $\bar{Y}$ even though the population is right-skewed.

6 Normal Distribution and Standardization

Question 6.1: Write down the standardization formula.

If $Y \sim N(\mu_Y, \sigma_Y^2)$, then the standardized variable:

$$Z = \frac{Y - \mu_Y}{\sigma_Y} \sim N(0,1)$$

Standardization subtracts the mean (centers at 0) and divides by the standard deviation (scales to unit variance).

Question 6.2: Suppose $\bar{Y} \overset{a}{\sim} N(700, 400)$, so $\mu_Y = 700$ and $\sigma_{\bar{Y}} = 20$.

(a) $\Pr(\bar{Y} \leq 740)$:

$$\Pr(\bar{Y} \leq 740) = \Pr\left(Z \leq \frac{740 - 700}{20}\right) = \Pr(Z \leq 2) = \boxed{0.9772}$$

(b) $\Pr(660 \leq \bar{Y} \leq 740)$:

$$\Pr(660 \leq \bar{Y} \leq 740) = \Pr\left(\frac{660 - 700}{20} \leq Z \leq \frac{740 - 700}{20}\right) = \Pr(-2 \leq Z \leq 2)$$

$$= \Phi(2) - \Phi(-2) = 0.9772 - 0.0228 = \boxed{0.9544}$$

(c) $\Pr(\bar{Y} > 730)$:

$$\Pr(\bar{Y} > 730) = 1 - \Pr\left(Z \leq \frac{730 - 700}{20}\right) = 1 - \Pr(Z \leq 1.5) = 1 - 0.9332 = \boxed{0.0668}$$

Question 6.3: What fraction falls within 1.96 standard errors?

By the CLT, $\bar{Y} \overset{a}{\sim} N(\mu_Y, \sigma_Y^2/n)$. Therefore:

$$\Pr\left(\mu_Y - 1.96 \cdot \frac{\sigma_Y}{\sqrt{n}} \leq \bar{Y} \leq \mu_Y + 1.96 \cdot \frac{\sigma_Y}{\sqrt{n}}\right) \approx 0.95$$

Approximately **95%** of sample means fall within 1.96 standard errors of the population mean. The simulation should confirm a fraction close to 0.95.

This result is the foundation for **confidence intervals** and **hypothesis testing** in econometrics.

7 Putting It All Together

Question 7.1: Explain the standardized CLT result in plain language.

The expression:

$$\frac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} \xrightarrow{d} N(0,1)$$

says that if we take the sample mean, subtract the population mean, and divide by its standard error, the result is approximately standard normal for large n.

Why it matters for econometrics:

- It allows us to make **probability statements** about how far $\bar{Y}$ might be from μ_Y
- It is the basis for **hypothesis testing**: we can test whether a claimed value of μ_Y is consistent with observed data
- It enables **confidence intervals**: we can construct ranges likely to contain μ_Y
- It works regardless of the population distribution, requiring only i.i.d. data and a large enough sample

Question 7.2: Applied problem with $n = 400$, $\bar{Y} = 712$, $\sigma_Y = 200$.

(a) **Standard error:**

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{200}{\sqrt{400}} = \frac{200}{20} = \boxed{10}$$

(b) By the CLT, $\bar{Y} \overset{a}{\sim} N(712, 100)$ under repeated sampling from this population. But the question asks about the probability under the true population mean. Since $\sigma_{\bar{Y}} = 10$:

$$\Pr(692 \leq \bar{Y} \leq 732) = \Pr\left(\frac{692 - \mu_Y}{10} \leq Z \leq \frac{732 - \mu_Y}{10}\right)$$

If we assume $\mu_Y = 712$:

$$= \Pr\left(\frac{-20}{10} \leq Z \leq \frac{20}{10}\right) = \Pr(-2 \leq Z \leq 2) = \boxed{0.9544}$$

(c) If the true mean were $\mu_Y = 750$, then the standardized value of our observed $\bar{Y} = 712$ would be:

$$Z = \frac{712 - 750}{10} = \frac{-38}{10} = -3.8$$

A Z-score of -3.8 is far in the left tail of the standard normal distribution. The probability of observing $\bar{Y} \leq 712$ if $\mu_Y = 750$ is:

$$\Pr(Z \leq -3.8) \approx 0.00007$$

This is extremely unlikely ($< 0.01\%$). Our sample provides strong evidence **against** the claim that $\mu_Y = 750$. The true mean is very likely lower than \$750.