Handout 3: Random Sampling & Large Sample Approximations

EC 282: Introduction to Econometrics

Spring 2026

Instructions: Run the provided R code and answer the questions. Show your work for calculations.

1 Population vs. Sample

Run the code below to create a population of weekly earnings for 100,000 workers.

```r
library(ggplot2)
set.seed(282)

# Create a population of weekly earnings (right-skewed)
N <- 100000
population <- data.frame(
  id = 1:N,
  earnings = round(500 + rgamma(N, shape = 2, rate = 0.01), 2)
)

pop_mean <- mean(population$earnings)
pop_var  <- var(population$earnings) * (N - 1) / N
pop_sd   <- sqrt(pop_var)

cat("Population mean (mu_Y):", pop_mean, "\n")
cat("Population variance (sigma_Y^2):", pop_var, "\n")
cat("Population std dev (sigma_Y):", pop_sd, "\n")
```

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

Question 1.2: Plot a histogram of the population earnings using the code below. Describe the shape of the distribution. Is it symmetric?

```r
ggplot(population, aes(x = earnings)) +
  geom_histogram(bins = 60, fill = "steelblue", color = "white") +
  geom_vline(xintercept = pop_mean, color = "red",
             linetype = "dashed", linewidth = 1) +
  labs(title = "Population Distribution of Weekly Earnings",
       x = "Weekly Earnings ($)", y = "Count") +
  theme_minimal()
```

2 Random Sampling

Question 2.1: What does it mean for a sample to be **i.i.d.** (independently and identically distributed)? Why is random sampling important for ensuring the i.i.d. property?

Question 2.2: Draw a random sample of $n = 10$ from the population and compute the sample mean. Then repeat for $n = 100$.

```r
sample_10  <- sample(population$earnings, 10)
sample_100 <- sample(population$earnings, 100)

cat("Sample mean (n = 10):", mean(sample_10), "\n")
cat("Sample mean (n = 100):", mean(sample_100), "\n")
cat("Population mean:", pop_mean, "\n")
```

(a) Are either of the sample means exactly equal to the population mean? Why or why not?

(b) Which sample mean is likely to be closer to the population mean? Explain.

3 The Sample Mean as a Random Variable

Question 3.1: Why is the sample mean $\bar{Y}$ a **random variable**? What would happen if you drew a different random sample of the same size?

Question 3.2: Using the definitions from your lecture notes, show mathematically that:

(a) $E[\bar{Y}] = \mu_Y$

(b) $\text{Var}(\bar{Y}) = \dfrac{\sigma_Y^2}{n}$

Hint: Use the fact that $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2$, and that for independent random variables, $\text{Var}(Y_i + Y_j) = \text{Var}(Y_i) + \text{Var}(Y_j)$.

Question 3.3: What is the standard error of $\bar{Y}$? Compute it for $n = 10$, $n = 100$, and $n = 1000$ using the population standard deviation.

```
for (n in c(10, 100, 1000)) {
  se <- pop_sd / sqrt(n)
  cat("n =", n, " -> Standard Error =", round(se, 4), "\n")
}
```

What happens to the standard error as n increases? Explain why this makes sense.

4 Law of Large Numbers

Question 4.1: State the Law of Large Numbers in your own words. What conditions must be satisfied for it to hold?

Question 4.2: Run the code below to see how the sample mean converges as n increases:

```
sample_sizes <- c(10, 25, 50, 100, 250, 500, 1000,
                  2500, 5000, 10000, 50000)

results <- data.frame(
  n = sample_sizes,
  sample_mean = sapply(sample_sizes, function(n) {
    mean(sample(population$earnings, n))
  })
)
results$pop_mean <- pop_mean
results$deviation <- results$sample_mean - pop_mean
print(results)
```

(a) Describe the pattern you observe in the deviation column.

(b) At what sample size does the sample mean start getting "very close" to the population mean?

Question 4.3: Run the code below to visualize the LLN with multiple samples at each size:

```
lln_data <- do.call(rbind, lapply(
  round(exp(seq(log(10), log(N), length.out = 80))),
  function(n) {
    data.frame(
      n = n,
      sample_mean = replicate(10,
        mean(sample(population$earnings, n)))
    )
  }
))

ggplot(lln_data, aes(x = n, y = sample_mean)) +
  geom_point(alpha = 0.3, color = "steelblue", size = 1.5) +
  geom_hline(yintercept = pop_mean, color = "red",
             linetype = "dashed", linewidth = 1) +
  scale_x_log10(labels = scales::comma) +
  labs(title = "Law of Large Numbers",
       subtitle = "Each dot is a sample mean; spread decreases with n",
       x = "Sample Size (log scale)", y = "Sample Mean") +
  theme_minimal()
```

Explain how this plot illustrates both the Law of Large Numbers and the formula $\text{Var}(\bar{Y}) = \sigma_Y^2/n$.

5 Central Limit Theorem

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

Question 5.2: Run the code below to draw 10,000 samples of size $n = 5$, $n = 30$, and $n = 200$ from the (skewed) earnings population:

```r
clt_data <- do.call(rbind, lapply(c(5, 30, 200), function(n) {
  data.frame(
    n = paste("n =", n),
    sample_mean = replicate(10000,
      mean(sample(population$earnings, n)))
  )
}))

ggplot(clt_data, aes(x = sample_mean)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "steelblue", color = "white", alpha = 0.7) +
  facet_wrap(~ n, scales = "free") +
  labs(title = "Sampling Distribution of the Mean",
       x = "Sample Mean", y = "Density") +
  theme_minimal()
```

(a) Describe the shape of the sampling distribution for each value of n.

(b) Recall that the population distribution is right-skewed. Why does the sampling distribution become symmetric even though the population is skewed?

(c) What happens to the spread (variance) of the sampling distribution as n increases?

Question 5.3: The CLT tells us that $\bar{Y} \overset{a}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$. Run the code below to overlay the normal approximation on the sampling distribution for $n = 100$:

```r
n <- 100
means_100 <- replicate(10000,
  mean(sample(population$earnings, n)))

ggplot(data.frame(x = means_100), aes(x = x)) +
  geom_histogram(aes(y = after_stat(density)), bins = 50,
                 fill = "steelblue", color = "white", alpha = 0.7) +
  stat_function(fun = dnorm,
    args = list(mean = pop_mean,
                sd = pop_sd / sqrt(n)),
    color = "darkred", linewidth = 1.2) +
  labs(title = "CLT: Normal Approximation (n = 100)",
       x = "Sample Mean", y = "Density") +
```

```
  theme_minimal()
```

How well does the normal curve fit the histogram? What does this tell you about the CLT approximation at $n = 100$?

6 Normal Distribution and Standardization

Question 6.1: If $Y \sim N(\mu_Y, \sigma_Y^2)$, write down the formula to **standardize** Y into a standard normal variable $Z \sim N(0, 1)$.

Question 6.2: Suppose the sampling distribution of $\bar{Y}$ is approximately $N(700, 400)$ (i.e., $\mu_Y = 700$ and $\sigma_{\bar{Y}}^2 = 400$, so $\sigma_{\bar{Y}} = 20$). Calculate:

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

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

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

Hint: Standardize $\bar{Y}$ and use the standard normal table or `pnorm()` in R.

```
# Verify your calculations in R
mu   <- 700
se   <- 20

# (a) P(Y_bar <= 740)
```

```r
pnorm(740, mean = mu, sd = se)

# (b) P(660 <= Y_bar <= 740)
pnorm(740, mean = mu, sd = se) - pnorm(660, mean = mu, sd = se)

# (c) P(Y_bar > 730)
1 - pnorm(730, mean = mu, sd = se)
```

Question 6.3: Using the CLT, approximately what fraction of sample means (with $n = 100$) fall within 1.96 standard errors of the population mean? Verify with the simulation:

```r
se_100 <- pop_sd / sqrt(100)
lower <- pop_mean - 1.96 * se_100
upper <- pop_mean + 1.96 * se_100

fraction_within <- mean(means_100 >= lower & means_100 <= upper)
cat("Fraction within 1.96 SE:", fraction_within, "\n")
cat("Theoretical:", 0.95, "\n")
```

7 Putting It All Together

Question 7.1: Explain in plain language what the following mathematical statement means and why it is important for econometrics:

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

Question 7.2: A survey of $n = 400$ workers finds that the average weekly earnings in the sample is $\bar{Y} = 712$ with a (known) population standard deviation of $\sigma_Y = 200$. Using the CLT:

(a) What is the standard error of $\bar{Y}$?

(b) What is the approximate probability that $\bar{Y}$ falls between \$692 and \$732? Show your standardization steps.

(c) A politician claims the true average weekly earnings is \$750. Based on our sample, does this seem plausible? Compute the standardized value and explain.