Handout 4: Estimation, Hypothesis Testing & Confidence Intervals

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Setup: The Data

Question 1.1: Compute $\bar{Y}$, S_Y, and $SE[\bar{Y}]$.

The sample mean:

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

The sample standard deviation:

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

The standard error:

$$SE[\bar{Y}] = \frac{S_Y}{\sqrt{n}}$$

Exact values depend on the random sample drawn. Run the R code to obtain the numerical values.

Question 1.2: Explain the difference between σ_Y and S_Y, and between $\sigma_{\bar{Y}}$ and $SE[\bar{Y}]$.

- σ_Y is the **population** standard deviation — a fixed but **unknown** parameter
- S_Y is the **sample** standard deviation — an **estimator** of σ_Y computed from data
- $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$ is the **true** standard deviation of the sampling distribution of $\bar{Y}$ — unknown because σ_Y is unknown
- $SE[\bar{Y}] = S_Y/\sqrt{n}$ is the **estimated** standard deviation of the sampling distribution — our best approximation using sample data

We need the sample-based versions because in practice we almost never know the population parameters σ_Y and $\sigma_{\bar{Y}}$. The sample versions are computable from data and, by the Law of Large Numbers, converge to the true values as n grows.

2 Estimators and Their Properties

Question 2.1: Estimator vs. estimate.

An **estimator** is a formula (a function of sample data) used to estimate an unknown population parameter. Since it depends on random data, an estimator is itself a **random variable**.

An **estimate** is the specific numerical value you get when you plug your actual sample data into the estimator. It is a **fixed number**.

In the rent example:

- **Estimator**: $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ (a formula — a random variable)
- **Estimate**: The computed value (e.g., $962.15 — a specific number from our particular sample)

Question 2.2: Three properties of good estimators.

(a) **Mathematical definitions:**

Unbiasedness: $E[\hat{\mu}_Y] = \mu_Y$

Consistency: $\hat{\mu}_Y \xrightarrow{p} \mu_Y$ as $n \to \infty$

Efficiency: Among all unbiased estimators, $\hat{\mu}_Y$ has the smallest variance: $\text{Var}(\hat{\mu}_Y) \leq \text{Var}(\tilde{\mu}_Y)$ for any other unbiased estimator $\tilde{\mu}_Y$.

(b) **Plain language:**

- **Unbiased**: On average (across many hypothetical samples), the estimator hits the true value. It does not systematically overestimate or underestimate.
- **Consistent**: As the sample size grows, the estimator gets closer and closer to the true value. With enough data, we can learn the truth.
- **Efficient**: Among estimators that are unbiased, this one has the least spread — its estimates are most tightly concentrated around the true value.

Question 2.3: Compare sample mean and sample median.

(a) **Both estimators are approximately unbiased.** The mean of the 10,000 simulated sample means should be very close to μ_Y, and the mean of the 10,000 simulated medians should also be close (since the population is normal and hence symmetric, the median is also unbiased).

(b) The sample mean has **smaller variance** than the sample median, so $\bar{Y}$ is **more efficient**. For a normal population, $\text{Var}(\text{median}) \approx \frac{\pi}{2} \times \text{Var}(\bar{Y}) \approx 1.57 \times \text{Var}(\bar{Y})$. This means the sample mean is the better estimator: it is BLUE (Best Linear Unbiased Estimator) of μ_Y.

Question 2.4: What property does the plot demonstrate?

The plot demonstrates **consistency**. As n increases from 10 to 5,000, the sampling distribution of $\bar{Y}$:

- Becomes **more concentrated** around μ_Y (narrower histograms)
- The variance shrinks ($\text{Var}(\bar{Y}) = \sigma_Y^2/n \to 0$)

By $n = 5{,}000$, nearly all sample means are very close to the red dashed line (the true μ_Y). This is $\bar{Y} \xrightarrow{p} \mu_Y$.

3 Hypothesis Testing

Question 3.1: Write the hypotheses.

Since the housing office suspects rents have *increased* beyond \$900, the natural setup is a **one-sided test**:

$$H_0 : \mu_Y = 900 \quad \text{(rents have not increased)}$$
$$H_A : \mu_Y > 900 \quad \text{(rents have increased)}$$

A **two-sided alternative** ($H_A : \mu_Y \neq 900$) would also be reasonable if the office simply wanted to know whether average rent differs from \$900 in either direction. We will compute p-values for both.

Question 3.2: Compute the t-statistic.

$$t\text{-stat} = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]} = \frac{\bar{Y} - 900}{S_Y/\sqrt{50}}$$

Plug in the values from R. For example, if $\bar{Y} = 962.15$ and $SE = 42.43$:

$$t\text{-stat} = \frac{962.15 - 900}{42.43} = 1.464$$

(Your exact numbers will come from the R output.)

Question 3.3: Compute the p-values.

(a) **Two-sided p-value** ($H_A : \mu_Y \neq 900$):

$$\text{p-value} = 2 \times \Phi(-|t\text{-stat}|)$$

This gives the probability of seeing a test statistic at least as extreme as ours in *either* direction, assuming H_0 is true. Verify with R.

(b) **One-sided p-value** ($H_A : \mu_Y > 900$):

$$\text{p-value} = 1 - \Phi(t\text{-stat})$$

This gives the probability of seeing a $\bar{Y}$ at least as large as ours if $\mu_Y = 900$. The one-sided p-value is exactly half the two-sided p-value (when $t > 0$).

Question 3.4: Decision at different significance levels.

Decision rule: Reject H_0 if p-value $< \alpha$.

- At $\alpha = 0.05$: Compare your p-value to 0.05. If p-value < 0.05, reject H_0; otherwise, fail to reject.

- At $\alpha = 0.01$: Compare your p-value to 0.01 (a stricter standard).

The decision depends on your actual sample. If, say, the two-sided p-value is 0.035, you would reject at $\alpha = 0.05$ but fail to reject at $\alpha = 0.01$. The stricter the significance level, the harder it is to reject H_0 (which reduces the chance of a Type I error).

Question 3.5: Visualize the rejection region.

The plot shows the standard normal distribution under H_0. The red shaded areas are the **rejection region** (both tails beyond ± 1.96 for $\alpha = 0.05$ two-sided). The blue dashed line marks our t-statistic.

- If the blue line falls **inside** the red region: we **reject** H_0
- If the blue line falls **outside** the red region: we **fail to reject** H_0

This should be consistent with whether the p-value is above or below 0.05.

4 Type I and Type II Errors

Question 4.1: Complete the table.

	Fail to Reject H_0	**Reject H_0**
H_0 True	Correct decision $(1-\alpha)$	**Type I Error** (α)
H_0 False	**Type II Error** (β)	Correct decision (Power $= 1-\beta$)

- **Type I Error** (false positive): Rejecting H_0 when it is actually true. Probability $= \alpha$.
- **Type II Error** (false negative): Failing to reject H_0 when it is actually false. Probability $= \beta$.
- **Power**: The probability of correctly rejecting H_0 when it is false. Power $= 1-\beta$.

Question 4.2: Type I and Type II errors in the rent example.

Type I Error: Concluding that average rent has increased beyond \$900 when in fact it has not. The housing office would unnecessarily increase the budget, wasting resources.

Type II Error: Concluding that average rent is still \$900 when in fact it has increased. The housing office would underfund housing support, leaving students struggling with higher rents.

Question 4.3: Type I error simulation.

The simulated Type I error rate should be very close to $\alpha = 0.05$ (typically between 0.045 and 0.055).

This makes sense because:

- We created a population where H_0 is **true** ($\mu_Y = 900$)
- We use the critical value 1.96, which corresponds to $\alpha = 0.05$
- By construction, we reject 5% of the time when H_0 is true
- This is exactly what the significance level α controls: the probability of a Type I error

If we used $\alpha = 0.01$ (critical value 2.576), the rejection rate would drop to about 1%.

5 Confidence Intervals

Question 5.1: Construct the 95% CI.

$$\begin{aligned} 95\%\ \text{CI} &= \bar{Y} \pm 1.96 \times SE[\bar{Y}] \\ &= \bar{Y} \pm 1.96 \times \frac{S_Y}{\sqrt{n}} \end{aligned}$$

Plug in the values from your R output. For example, if $\bar{Y} = 962.15$ and $SE = 42.43$:

$$95\%\ \text{CI} = 962.15 \pm 1.96 \times 42.43 = 962.15 \pm 83.16 = \boxed{[878.99,\ 1045.31]}$$

(Your exact numbers will come from R.)

Question 5.2: Compare 90%, 95%, and 99% CIs.

The three intervals use different critical values:

- 90% CI: $\bar{Y} \pm 1.65 \times SE$ (narrowest)
- 95% CI: $\bar{Y} \pm 1.96 \times SE$ (middle)
- 99% CI: $\bar{Y} \pm 2.576 \times SE$ (widest)

Trade-off: Higher confidence requires a wider interval. To be "more sure" that we capture μ_Y, we must cast a wider net. The 99% CI is the widest (most conservative but least precise), and the 90% CI is the narrowest (most precise but less confident).

Question 5.3: Correct interpretation.

The correct answer is **(b)**: "If we drew many samples and computed a 95% CI from each, about 95% of those intervals would contain the true μ_Y."

Why (a) is wrong: μ_Y is a fixed (non-random) number. It either is or is not in our interval. There is no probability involved for μ_Y itself. The randomness comes from *which interval we happen to compute.*

Why (c) is wrong: The CI is about the *population mean*, not about individual observations. Individual rents have much more spread than the sampling distribution of $\bar{Y}$.

Question 5.4: Coverage simulation.

Out of 100 intervals, approximately **95** should contain the true μ_Y (colored blue), and about **5** should miss it (colored red). The coverage rate should be close to 0.95.

This directly illustrates the correct interpretation: the "95%" refers to the **procedure's long-run success rate**, not to any single interval. If you ran this simulation 1,000 times, the average coverage would converge to exactly 0.95.

6 The Connection Between CIs and Hypothesis Tests

Question 6.1: Using the CI to test $H_0 : \mu_Y = 900$.

Yes! A two-sided hypothesis test at significance level α is equivalent to checking whether the hypothesized value falls inside the $(1 - \alpha) \times 100\%$ confidence interval:

- If $\mu_{Y,0} = 900$ falls **outside** the 95% CI $\Rightarrow$ **reject** H_0 at $\alpha = 0.05$
- If $\mu_{Y,0} = 900$ falls **inside** the 95% CI $\Rightarrow$ **fail to reject** H_0 at $\alpha = 0.05$

Check whether 900 lies within your CI bounds from Question 5.1.

Question 6.2: Test $H_0 : \mu_Y = 950$.

Since 950 is likely **inside** the 95% CI (it is close to $\bar{Y}$), we would **fail to reject** $H_0 : \mu_Y = 950$. The t-statistic should be small (close to 0) and the p-value should be large (well above 0.05). Verify with R.

General rule:

$$\boxed{\text{Reject } H_0 : \mu_Y = \mu_{Y,0} \text{ at } \alpha = 0.05 \quad \Longleftrightarrow \quad \mu_{Y,0} \notin 95\% \text{ CI}}$$

The 95% CI is exactly the set of all values $\mu_{Y,0}$ that we would *not* reject at the 5% level. This duality means CIs and hypothesis tests always give the same answer.

7 Comparing Two Groups

Question 7.1: Hypotheses.

Let μ_1 = average rent for off-campus students and μ_2 = average rent for university-adjacent students.

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{(no difference in rents)}$$
$$H_A : \mu_1 - \mu_2 \neq 0 \quad \text{(rents differ between the two groups)}$$

Question 7.2: t-statistic and p-value.

The standard error for the difference in means:

$$SE[\bar{Y}_1 - \bar{Y}_2] = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

The t-statistic:

$$t\text{-stat} = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{SE[\bar{Y}_1 - \bar{Y}_2]}$$

The two-sided p-value:

$$\text{p-value} = 2 \times \Phi(-|t\text{-stat}|)$$

Plug in the values from R. Since the true difference is $1020 - 950 = 70$, the t-statistic should be positive and fairly large, yielding a small p-value.

Question 7.3: 95% CI for the difference.

$$95\%\ \text{CI} = (\bar{Y}_1 - \bar{Y}_2) \pm 1.96 \times SE[\bar{Y}_1 - \bar{Y}_2]$$

If the interval does **not contain 0**, this is consistent with **rejecting** H_0 (same CI–test duality as before). If it **does contain 0**, this is consistent with **failing to reject** H_0.

Question 7.4: Interpretation.

Based on our samples, off-campus students pay significantly more in rent than students in university-adjacent housing. The difference in sample means is positive, and the 95% CI for $\mu_1 - \mu_2$ likely does not contain zero, confirming the statistical significance.

However, we should be cautious:

- Statistical significance does not imply that location *causes* higher rent — students who choose off-campus housing may differ in other ways (preferences for larger spaces, different neighborhoods, etc.)
- The CI tells us the plausible *range* of the difference, which is useful for budgeting purposes
- The housing office might use this evidence to adjust support for off-campus students or

to investigate why the gap exists