ISE 315: Engineering Statistics

Homework 1 Solution

Class Section: F-04, Instructor: Mansur M. Arief
Department of Industrial and Systems Engineering, KFUPM

Problem 1 [30 pt] — Comparing Estimators for Proportions

Context: LLM harmful response rate $p = 0.08$, batch $n = 50$. Four estimators compared via 5,000 simulations.

Part (a) [10 pt]: Classify Bias, Variance, and MSE

Estimator	Bias	Variance	MSE
Sample Proportion ($\hat{p}$)	Low (mean $\approx p$; unbiased by construction)	Medium	Low (0.00150)
Laplace Smoothing ($\hat{p}_L$)	Low (slight positive shift from $+1/+2$)	Low (tightest spread)	Low (0.00165)
Reduced Sample ($\hat{p}_R$)	Low (still unbiased; uses $X/10$)	High (widest violin)	High (0.00733)
Conservative ($\hat{p}_C$)	High (mean visibly above p; adds $+2/n$)	Medium (same as $\hat{p}$)	Medium (0.0031

Key observations from the figure:

- The violin for Reduced Sample is much wider than the others, confirming high variance.
- Conservative's diamond marker sits clearly above the true $p = 0.08$ line, indicating high bias.
- Sample Proportion and Laplace are nearly centered on p, but Laplace is slightly narrower.

Rubric — Part (a)

Error	Deduction
Correct classification but missing justification from the figure	−3
Incorrect bias classification (e.g., calling Conservative "Low")	−1 each
Incorrect variance or MSE classification	−1 each
Missing or no work shown	−8

Part (b) [8 pt]: MSE Decomposition Analysis

1. **Lowest MSE:** The **Sample Proportion** has the lowest MSE (0.00150). Although Laplace Smoothing has lower variance (0.00148 vs. 0.00150), its small bias ($\hat{p}_L$ systematically shifts toward 1/2 due to the $+1/+2$ terms) adds a Bias^2 component, bringing its total MSE to 0.00165. The Sample Proportion, being exactly unbiased, has MSE = Var + 0 = 0.00150.

Note: Accept Laplace as "lowest MSE" if the student correctly argues from the bar chart—the values are close.

2. **Conservative vs. Sample Proportion:** Both share the same variance because $\hat{p}_C = (X + 2)/n$ and $\hat{p} = X/n$ differ only by the constant $2/n = 0.04$. Adding a constant shifts every estimate but does not change the spread. However, this shift creates a large bias: $E[\hat{p}_C] = p + 0.04 = 0.12 \neq 0.08$, so $\text{Bias}^2 = 0.04^2 = 0.0016$. This bias^2 nearly *doubles* the MSE compared to the Sample Proportion.

Why This Matters

This illustrates the MSE decomposition vividly: adding a constant to an estimator **preserves variance** but **introduces bias**. The MSE penalty is Bias^2, which grows quadratically with the shift. We will see this same principle when studying confidence interval width (HW2)—centering matters.

Rubric — Part (b)

Error	Deduction
Part 1: Correct identification but no explanation of bias-variance trade-off	−2
Part 2: Does not explain why variances are equal (constant shift)	−2
Part 2: Does not quantify or explain how bias inflates MSE	−2
Missing or no work shown	−6

Part (c) [6 pt]: Safety-Critical Estimator Choice

(i) **Recommend: Conservative estimator** ($\hat{p}_C$). It systematically overestimates p (bias > 0), so it is least likely to underestimate the true harmful rate. In AI safety, the cost of deploying an unsafe system (underestimation) far exceeds the cost of extra review (overestimation). The Conservative estimator builds in a safety margin by design.

(ii) **Trade-off:** Higher MSE and more false alarms. The system will be flagged for review more often than necessary (higher false positive rate), wasting engineering resources on investigations that find nothing wrong. There is also the risk of "alarm fatigue"—if the team is constantly reviewing false positives, they may become desensitized to real alerts.

Rubric — Part (c)

Error	Deduction
Part (i): Correct choice but no justification tied to overestimation	−1
Part (i): Wrong choice (e.g., recommends Sample Proportion)	−2
Part (ii): Does not mention false alarms or resource cost	−2
Missing or no work shown	−5

Part (d) [6 pt]: Why Reduced Sample Has High Variance

(i) **High variance:** The Reduced Sample uses only $n = 10$ out of 50 available observations. Since $\text{Var}(\hat{p}_R) = p(1-p)/10$ versus $\text{Var}(\hat{p}) = p(1-p)/50$, the variance is $5\times$ larger. Discarding 80% of the data throws away information, directly inflating the standard error by a factor of $\sqrt{5} \approx 2.24$.

(ii) **Practical justification:** You might prefer this estimator when the full batch of 50 responses is **not yet available** and a quick decision is needed. For example, in real-time monitoring, if the first 10 responses already show a high harmful rate, waiting for all 50 would delay intervention. Another scenario: if evaluating each response for harm is expensive (e.g., requires human review), using only 10 saves significant cost while still providing an unbiased estimate. The trade-off is precision for speed (or cost savings).

Rubric — Part (d)

Error	Deduction
Part (i): Does not connect smaller n to larger variance formula	-1
Part (ii): Gives a scenario but does not explain the trade-off	-2
Part (ii): No practical scenario given	-3
Missing or no work shown	-5

Problem 2 [30 pt] — CLT with Skewed Distributions

Context: LLM response time $X \sim \text{Exponential}(\lambda = 0.5)$, so $\mu = \sigma = 1/\lambda = 2.0$ seconds.

Part (a) [10 pt]: Comparison Table

Property	**Population (Individual X)**	**Sampling Dist. of $\bar{X}$**		
		$n = 2$	$n = 10$	$n = 50$
Shape	Highly right-skewed	Right-skewed	Mildly skewed	Approx. Normal
Mean	$\mu = 2.0$ s	Same: 2.0	Same: 2.0	Same: 2.0
Std. Dev. (or SE)	$\sigma = 2.0$ s	$\frac{2.0}{\sqrt{2}} = 1.41$	$\frac{2.0}{\sqrt{10}} = 0.63$	$\frac{2.0}{\sqrt{50}} = 0.29$
Skewness (approx.)	2.0 (high)	1.41	0.60	0.32

Why the mean stays the same: $E[\bar{X}] = \mu$ for any n—the sample mean is always unbiased.
Why SE and skewness both shrink: Averaging n observations compresses the spread by $1/\sqrt{n}$ and, per the CLT, progressively "normalizes" the shape. The skewness of $\bar{X}$ for an Exponential population is $2/\sqrt{n}$, which decreases as n grows.

Rubric — Part (a)

Error	**Deduction**
Missing specific numerical SE values (only wrote "decreases")	−3
Incorrect shape descriptions (e.g., calling $n = 2$ "Normal")	−2
Incorrect SE computation	−2
No explanation for why properties change or stay constant	−2
Missing or no work shown	−8

Part (b) [8 pt]: Correcting the Misconception

(i) **Why incorrect:** The Central Limit Theorem guarantees that the sampling distribution of $\bar{X}$ approaches a Normal distribution as n increases, **regardless** of the population shape. The key condition is that $\bar{X}$ is a mean (normalized sum) of i.i.d. observations with finite variance. Even though individual response times are highly right-skewed (skewness = 2.0), the skewness of $\bar{X}$ decreases as $2/\sqrt{n}$. By $n = 50$, the skewness is only 0.32 and the histogram closely tracks the Normal approximation curve.

(ii) **Best Normal approximation:** $n = 50$. Visual evidence: the simulated histogram (blue bars) at $n = 50$ is nearly symmetric and fits the dashed Normal curve almost perfectly. The skewness has dropped to 0.32 (close to 0), and there is no visible right tail excess. By contrast, at $n = 2$ the histogram is still clearly right-skewed with visible departure from the Normal curve.

Rubric — Part (b)

Error	Deduction
Part (i): Mentions CLT but does not specify it applies to the *sample mean*	−1
Part (i): Says CLT applies to "all statistics" generically	−2
Part (i): No reference to CLT at all	−3
Part (ii): Wrong sample size or no visual evidence cited	−2
Missing or no work shown	−6

Part (c) [12 pt]: SLA Monitoring Analysis

(i) $n = 10$, **observing** $\bar{X} = 3.5$ **s:** Under normal operation ($\mu = 2.0$, $\sigma = 2.0$):

$$\text{SE} = \frac{2.0}{\sqrt{10}} \approx 0.632, \qquad Z = \frac{3.5 - 2.0}{0.632} \approx 2.37.$$

From the Z-table: $P(\bar{X} > 3.5) = P(Z > 2.37) \approx 0.009$, i.e., less than 1%.

Yes, concerned. An observation this extreme would occur less than 1% of the time under normal operation. This is strong evidence that the mean response time may have increased. However, from Figure 3 (variability comparison), the 90% interval for sample means at $n = 10$ is roughly $[1.1, 3.2]$ seconds, and 3.5 falls outside this range—consistent with our Z-calculation.

(ii) $n = 30$, **observing** $\bar{X} = 3.5$ **s:**

$$\text{SE} = \frac{2.0}{\sqrt{30}} \approx 0.365, \qquad Z = \frac{3.5 - 2.0}{0.365} \approx 4.11.$$

$P(Z > 4.11)$ is essentially 0 (< 0.00002).

Conclusion: With the larger sample, the same $\bar{X} = 3.5$ is now overwhelmingly unlikely under $\mu = 2.0$. The manager should be highly confident that the true mean response time has increased and should investigate the cause (e.g., model overload, infrastructure issues). The tripled sample size reduced the SE, making the test far more sensitive to deviations from μ.

Why This Matters

The jump from $Z = 2.37$ to $Z = 4.11$ by merely tripling the sample size illustrates how **power increases with** n. The same observed value becomes "more surprising" under H_0 when we have more data. This is exactly what drives sample size planning in hypothesis testing (Chapter 9): larger n means smaller SE, which means we can detect smaller departures from the null.

Rubric — Part (c)

Error	Deduction
Part (i): Uses σ instead of $\sigma/\sqrt{n}$	-3
Part (i): Correct calculation but no interpretation of whether to be concerned	-2
Part (ii): Does not compare to part (i) or explain the effect of larger n	-3
Arithmetic errors in SE or Z	-2 each
Missing or no work shown	-10

Problem 3 [40 pt] — LLM Content Moderation Monitoring

Context: Harmful response proportion $p = 0.05$, batch size $n = 200$, sample proportion $\hat{p} = X/n$.

Part (a) [10 pt]: Basic Calculations

(i) **Expected value and SE:**

$$E[\hat{p}] = p = 0.05, \qquad \text{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.05 \times 0.95}{200}} = \sqrt{\frac{0.0475}{200}} = \sqrt{0.0002375} \approx 0.0154.$$

(ii) $P(\hat{p} > 0.07)$**:** Standardize using $Z = \dfrac{\hat{p} - p}{\text{SE}}$:

$$Z = \frac{0.07 - 0.05}{0.0154} = \frac{0.02}{0.0154} \approx 1.30.$$

$$P(\hat{p} > 0.07) = P(Z > 1.30) = 1 - \Phi(1.30) = 1 - 0.9032 \approx 0.0968.$$

(iii) **Unusual?** At ∼9.7%, this is somewhat uncommon but not extremely rare. It would not meet the conventional 5% threshold for strong evidence of a problem. One might flag it for monitoring but not for immediate action.

$$\boxed{E[\hat{p}] = 0.05, \quad \text{SE} \approx 0.0154, \quad P(\hat{p} > 0.07) \approx 0.097}$$

Rubric — Part (a)

Error	Deduction
Wrong SE formula (e.g., missing square root, uses p^2 instead of $p(1-p)$)	−3
Correct SE but wrong Z-value	−2
Correct calculation but no interpretation of whether it's unusual	−2
Missing steps	−3
Missing or no work shown	−8

Part (b) [10 pt]: False Alarm Probability

Find: $P(\hat{p} < 0.03 \text{ or } \hat{p} > 0.07)$ under $p = 0.05$, with SE $= 0.0154$.

Step 1: Standardize both bounds:

$$Z_{\text{lower}} = \frac{0.03 - 0.05}{0.0154} = \frac{-0.02}{0.0154} \approx -1.30, \qquad Z_{\text{upper}} = \frac{0.07 - 0.05}{0.0154} \approx 1.30.$$

Step 2: By symmetry of the interval around p:

$$P(\text{false alarm}) = P(|Z| > 1.30) = 2 \times P(Z > 1.30) = 2 \times 0.0968 \approx 0.1936.$$

$$\boxed{P(\text{false alarm}) \approx 0.194}$$

Interpretation: About 19.4% of the time, a properly functioning system would be incorrectly flagged. This is a high false alarm rate—the thresholds $[0.03, 0.07]$ are not very stringent.

Why This Matters

Compare this to Problem 3 in F-03's homework, where the threshold $\mu \pm 2\,\text{SE}$ gave a false alarm rate of $\sim$4.6%. Here, $0.07 - 0.05 = 0.02 \approx 1.30\,\text{SE}$ (not $2\,\text{SE}$), so the interval is narrower in SE-units, hence more false alarms. When we formalize this as hypothesis testing, this false alarm rate is exactly the **significance level** α.

Rubric — Part (b)

Error	**Deduction**
Uses one-sided instead of two-sided probability	-4
Correct Z-values but wrong probability computation	-3
Uses wrong SE (e.g., from wrong p or wrong formula)	-4
Missing or no work shown	-8

Part (c) [10 pt]: Minimum Batch Size + Practical Trade-off

Part (i) [5 pt]: Minimum n.

Step 1: Set up: $\text{SE}(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}} \leq 0.01$.

Step 2: Square both sides:

$$\frac{0.05 \times 0.95}{n} \leq 0.01^2 = 0.0001 \quad \Longrightarrow \quad n \geq \frac{0.0475}{0.0001} = 475.$$

$$\boxed{n \geq 475}$$

Part (ii) [5 pt]: Practical trade-off.

At 50 responses per minute, collecting $n = 475$ responses takes $475/50 = 9.5$ minutes. This means the monitoring system would need to wait nearly 10 minutes before it can make a detection decision with SE ≤ 0.01.

Trade-off: This is the **precision vs. detection speed** dilemma in real-time content moderation. A smaller batch ($n = 200$, SE $= 0.0154$) can detect in 4 minutes but with coarser resolution—it might miss a moderate increase in harmful content. A larger batch ($n = 475$, SE $= 0.01$) can detect smaller shifts but introduces a ∼10-minute blind spot during which harmful content may propagate unchecked. In practice, engineers might use a *sequential monitoring* approach: start with quick, rough checks and escalate to larger batches only when initial signals are ambiguous.

Rubric — Part (c)

Error	Deduction
Wrong SE formula or incorrect algebra	−3
Does not round up (if a non-integer arose)	−1
Correct n but missing engineering discussion entirely	−5
Discussion present but does not compute collection time	−2
Does not articulate precision vs. speed trade-off	−2
Missing or no work shown	−8

Part (d) [10 pt]: Detection Power Under Attack

Given: True proportion shifts to $p' = 0.10$, $n = 200$, threshold $= 0.07$.

Part (i) [8 pt]: Power calculation.

Step 1: Under the degraded model: $\text{SE}_{\text{alt}} = \sqrt{\dfrac{0.10 \times 0.90}{200}} = \sqrt{\dfrac{0.09}{200}} = \sqrt{0.00045} \approx 0.0212$.

Step 2: Standardize the threshold *under the new proportion*:

$$Z = \frac{0.07 - 0.10}{0.0212} = \frac{-0.03}{0.0212} \approx -1.41.$$

Step 3: Look up:

$$P(\hat{p} > 0.07) = P(Z > -1.41) = \Phi(1.41) \approx 0.9207.$$

$$\boxed{\text{Power} = P(\text{detection}) \approx 0.921}$$

The system correctly detects the attack about 92% of the time—good but not excellent (8% of attacks would be missed).

Part (ii) [2 pt]: How to increase power.

Increase the batch size n. Larger n reduces the SE under both the null and alternative distributions, making the threshold easier to exceed when the true proportion has shifted. For example, doubling to $n = 400$ would reduce SE_{alt} to ≈ 0.015, giving $Z = (0.07 - 0.10)/0.015 = -2.0$ and power ≈ 0.977.

Other acceptable answers: Use a lower threshold (e.g., 0.065 instead of 0.07), though this increases the false alarm rate.

Why This Matters

Part (b) gave us $\alpha \approx 0.194$ (Type I error / false alarm rate) and Part (d) gives power ≈ 0.921 (i.e., $\beta \approx 0.079$ Type II error). These are **not** complementary ($\alpha + \beta \neq 1$) because they are computed under different distributions—null vs. alternative. This asymmetry is fundamental to hypothesis testing (Chapter 9), where we will formalize the trade-off between α and power as a function of n, effect size, and threshold.

Rubric — Part (d)

Error	Deduction
Uses null $p = 0.05$ instead of $p' = 0.10$ for SE	-4
Correct SE but wrong standardization or arithmetic	-3
Does not use the alternative p' in variance formula	-3
Part (ii): No concrete suggestion or vague answer	-1
Missing or no work shown	-8

Total: $30 + 30 + 40 =$ **100 points.**