Chapter 5- Statistical Inference and Analysis

Question 1

A data scientist collects a sample of 80 customer transactions from an online store. The sample mean transaction value is €156.40 and the sample standard deviation is €42.30.

a. What is the point estimate for the population mean transaction value?

b. Calculate the 95% confidence interval for the true average transaction value. Show all steps including the appropriate t-value.

c. Interpret this confidence interval in plain language that a non-statistician could understand.

d. If you wanted a narrower confidence interval, what three changes could you make? Explain briefly.

Solution

Given:

- Sample size $n = 80$
- Sample mean $\bar{x} = €156.40$
- Sample standard deviation $s = €42.30$
- Confidence level = 95%

a. Point Estimate

The point estimate for the population mean transaction value is the sample mean:

$$\hat{\mu} = \bar{x} = €156.40$$

b. 95% Confidence Interval:

Since σ is unkown and we have a sample, we use the t-distribution.

- **Step 1:** Find degrees of freedom (df):
  - $df = n - 1 = 80 - 1 = 79$
- **Step 2:** Find the critical t-value
  - For a 95% CI: $\alpha = 0.05$, so $\alpha/2 = 0.025$
  - $t(0.025, 79) \approx 1.990$ (from t-distribution table)

- **Step 3:** Calculate the standard error (SE):
  - $SE = \frac{s}{\sqrt{n}} = \frac{42.30}{\sqrt{80}} \approx 4.73$
- **Step 4:** Calculate the margin of error (ME):
  - $ME = t_{\alpha/2} \times SE = 1.990 \times 4.73 \approx 9.41$
- **Step 5:** Construct the interval:
  - $CI = \bar{x} \pm ME = 156.40 \pm 9.41$

Thus, the 95% confidence interval is approximately $[€146.99, €165.81]$

c. We are 95% confident that the true average transaction value for all customers lies between €146.99 and €165.81. This means that if we repeated this sampling process many times and constructed a 95% confidence interval each time, approximately 95% of those intervals would contain the true population mean.

d. To obtain a narrower confidence interval, you could:

- **Increase sample size (n)** - reduces standard error.
- **Reduce confidence level** ***(e.g., from 95% to 90%)*** - smaller critical value.
- **Reduce variability in data** - though this is not directly controllable

Question 2

A machine learning model is tested on 400 samples and correctly classifies 348 of them.

a. Calculate the point estimate for the model's true accuracy.

b. Compute the 95% confidence interval for the true accuracy using the formula for proportions.

c. Based on this interval, can you conclude with 95% confidence that the model's true accuracy exceeds 85%? Justify your answer.

Solution

Given:

- Total samples $n = 400$
- Correct classifications $x = 348$

a. Point Estimate:

The point estimate for the model's true accuracy is the sample proportion of correct classifications:

$$\hat{p} = \frac{x}{n} = \frac{348}{400} = 0.87 \quad \text{or } 87\%$$

b. 95% Confidence Interval for Proportion:

Formula: $p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$

- **Step 1:** Find the critical z-value
  - For a 95% CI: $\alpha = 0.05$, so $\alpha/2 = 0.025$
  - $z(0.025) \approx 1.96$ (from z-table)
- **Step 2:** Calculate the standard error (SE):
  - $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.87(1-0.87)}{400}} \approx 0.0163$
- **Step 3:** Calculate the margin of error (ME):
  - $ME = z_{\alpha/2} \times SE = 1.96 \times 0.0163 \approx 0.0319$
- **Step 4:** Construct the interval:
  - $CI = \hat{p} \pm ME = 0.87 \pm 0.0319$

Thus, the 95% confidence interval is approximately $[0.8381, 0.9019]$ or $[83.81\%, 90.19\%]$.

c. Since the lower bound of the 95% confidence interval is approximately 83.81% , which is less than 85% , we cannot conclude with 95% confidence that the model's true accuracy exceeds 85% .

Question 3

You are given the following information about a sample:

- Sample size: $n = 25$
- Sample mean: $\bar{x} = 47.8$
- Population standard deviation: $\sigma = 8.5$ (known)

a. Would you use a z-distribution or t-distribution for constructing a confidence interval? Explain why.

b. Construct a 99% confidence interval for the population mean.

c. How would your answer to part (a) change if σ were unknown?

Solution

Given:

- Sample size $n = 25$
- Sample mean $\bar{x} = 47.8$

- Population standard deviation $\sigma = 8.5$ (known)

a. Distribution Choice:

We use the **z-distribution**.

Reasoning: The population standard deviation (σ) is known. When σ is known, we use the z-distribution regardless of sample size. The t-distribution is only used when σ is unknown and we must estimate it with the sample standard deviation s .

b. 99% Confidence Interval:

Formula: $\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$

- **Step 1:** Find the critical z-value
  - For a 99% CI: $\alpha = 0.01$, so $\alpha/2 = 0.005$
  - $z(0.005) = 2.576$ (from z-table)
- **Step 2:** Calculate the standard error (SE):
  - $SE = \frac{\sigma}{\sqrt{n}} = \frac{8.5}{\sqrt{25}} = \frac{8.5}{5} = 1.7$
- **Step 3:** Calculate the margin of error (ME):
  - $ME = z_{\alpha/2} \times SE = 2.576 \times 1.7 \approx 4.379$
- **Step 4:** Construct the interval:
  - $CI = \bar{x} \pm ME = 47.8 \pm 4.379$

Thus, the 99% confidence interval is approximately $[43.42, 52.18]$.

c. If σ were unknown, we would:

- Use the **t-distribution** instead of the z-distribution
- Use the sample standard deviation (s) instead of σ
- Use $t(0.005, 24) \approx 2.797$ instead of $z = 2.576$
- This would result in a **wider confidence interval**

Question 4

A researcher computes a 90% confidence interval for a population mean and obtains $[23.5, 28.7]$.

a. What is the sample mean?

b. What is the margin of error?

c. If the researcher wanted to construct a 95% confidence interval instead (keeping everything else the same), would the new interval be wider or narrower? Explain.

d. True or False: "There is a 90% probability that the true population mean lies between 23.5 and 28.7." If false, provide the correct interpretation.

Solution

Given: 90% CI = $[23.5, 28.7]$

a. Sample Mean:

The sample mean is the midpoint of the confidence interval:

$$\bar{x} = \frac{\text{lower bound} + \text{upper bound}}{2} = \frac{23.5 + 28.7}{2} = \frac{52.2}{2} = 26.1$$

b. Margin of Error:

The margin of error is half the width of the confidence interval:

$$ME = \frac{\text{upper bound} - \text{lower bound}}{2} = \frac{28.7 - 23.5}{2} = \frac{5.2}{2} = 2.6$$

c. 95% CI - Wider or Narrower?

The 95% confidence interval would be **WIDER**.

Explanation: A higher confidence level requires a larger critical value (t or z). For example:

- 90% CI uses $t(0.05, df)$ or $z = 1.645$
- 95% CI uses $t(0.025, df)$ or $z = 1.96$

Since everything else remains constant, the larger multiplier creates a larger margin of error and thus a wider interval.

d. True or False:

FALSE

The statement is incorrect because the confidence interval is about the procedure, not about probability after the interval is calculated.

Correct interpretation: "We are 90% confident that the true population mean lies between 23.5 and 28.7. This means that if we repeated this sampling process many times and constructed 90% confidence intervals, approximately 90% of those intervals would contain the true mean. Once the interval is constructed, the true mean either is or isn't in it—there's no probability involved at that point."

Question 5

A streaming platform claims that their new recommendation algorithm increases the average watch time per user from 145 minutes per week. A sample of 100 users using the new algorithm shows a mean watch time of 152 minutes with a standard deviation of 28 minutes. Test at $\alpha = 0.05$.

a. State the null and alternative hypotheses using proper notation. Is this a one-tailed or two-tailed test?

b. Calculate the test statistic.

c. Find the critical value(s) for this test.

d. What is your decision? State your conclusion in the context of the problem.

Solution

Given:

- Claimed mean: $\mu_0 = 145$ minutes
- Sample size: $n = 100$
- Sample mean: $\bar{x} = 152$ minutes
- Sample standard deviation: $s = 28$ minutes
- Significance level: $\alpha = 0.05$

a. Hypotheses:

$$H_0 : \mu \leq 145 \quad \text{(watch time has not increased)}$$

$$H_1 : \mu > 145 \quad \text{(watch time has increased)}$$

This is a **one-tailed (right-tailed) test** because we're only interested in whether watch time has increased.

b. Test Statistic:

Formula: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$

$$t = \frac{152 - 145}{28/\sqrt{100}} = \frac{7}{28/10} = \frac{7}{2.8} = 2.50$$

c. Critical Value:

- Degrees of freedom: $df = n - 1 = 100 - 1 = 99$
- For a one-tailed test with $\alpha = 0.05$:
  - $t(0.05, 99) \approx 1.660$
  - (For large df, this approaches $z = 1.645$)

d. Decision and Conclusion:

Decision Rule: Reject H_0 if $t > t_{\text{critical}}$

Since $2.50 > 1.660$, we **REJECT** H_0 .

Conclusion: At the 0.05 significance level, there is sufficient statistical evidence to conclude that the new recommendation algorithm increases the average watch time per user beyond 145 minutes per week. The observed increase from 145 to 152 minutes is statistically significant and unlikely to have occurred by chance alone if the true mean were still 145 minutes or less.

Question 6

An A/B test is conducted on a website:

- **Control group (A)**: 800 visitors, 64 conversions
- **Treatment group (B)**: 800 visitors, 88 conversions

The hypothesis test yields a p-value of 0.018.

a. State appropriate null and alternative hypotheses for this test.

b. Using $\alpha = 0.05$, what decision would you make?

c. Interpret the p-value in plain language.

d. If the p-value had been 0.073 instead, what would your conclusion be? Does this mean the treatment has no effect? Explain carefully.

Solution

Given:

- Control (A): $n_1 = 800$, conversions $= 64$, $\hat{p}_1 = 0.08$
- Treatment (B): $n_2 = 800$, conversions $= 88$, $\hat{p}_2 = 0.11$
- p-value $= 0.018$
- $\alpha = 0.05$

a. Hypotheses:

$$H_0 : p_B \leq p_A \quad \text{or} \quad p_B - p_A \leq 0$$

$$\text{(Treatment does not increase conversion rate)}$$

$$H_1 : p_B > p_A \quad \text{or} \quad p_B - p_A > 0$$

$$\text{(Treatment increases conversion rate)}$$

b. Decision:

Since p-value $(0.018) < \alpha(0.05)$, we **REJECT** H_0 .

c. Interpret p-value:

The p-value of 0.018 means: "If the treatment truly had no effect on conversion rate (if H_0 were true), there would be only a 1.8% chance of observing a difference in conversion rates as large as (or larger than) what we observed (11% vs 8%), purely due to random sampling variation. Since this probability is very small, it suggests our observed difference is unlikely to be due to chance alone."

d. If p-value = 0.073:

Since $0.073 > 0.05$, we would **FAIL TO REJECT** H_0.

Conclusion: We would not have sufficient statistical evidence at the 0.05 significance level to conclude that the treatment increases conversion rates.

Does this mean no effect? No, it does NOT mean the treatment has no effect. It means:

- The data do not provide strong enough evidence to claim an effect exists
- The observed difference (3 percentage points) could reasonably be explained by random sampling variation
- We haven't proven the null hypothesis true; we simply lack sufficient evidence to reject it
- There might be a real effect that we failed to detect (Type II error), perhaps due to insufficient sample size

Question 7

Consider the following statements about p-values. Indicate whether each is TRUE or FALSE, and briefly explain why.

a. A p-value of 0.03 means there is a 3% chance that the null hypothesis is true.

b. If $p = 0.48$, we accept the null hypothesis.

c. A very large sample size can produce a statistically significant result ($p < 0.05$) even when the actual difference is practically meaningless.

d. The p-value is calculated assuming the alternative hypothesis is true.

Solution

a. FALSE

The p-value is NOT the probability that H_0 is true.

The p-value is calculated **assuming** H_0 is true. It represents the probability of observing data as extreme as (or more extreme than) what we observed, given that the null hypothesis is true. The null hypothesis is either true or false—we don't assign probabilities to it after the fact.

b. FALSE

We never "accept" the null hypothesis. We either "reject H_0 " or "fail to reject H_0 ."

When $p = 0.48 > 0.05$, we fail to reject H_0 , meaning we don't have sufficient evidence against it. This is different from proving it true or "accepting" it. The data are simply consistent with H_0 , but that doesn't prove H_0 is correct.

c. TRUE

With very large samples, even tiny, practically meaningless differences can become statistically significant. This is because:

- Standard error decreases as n increases: $SE = s/\sqrt{n}$
- Smaller SE means larger test statistics for the same observed difference
- Statistical significance measures whether an effect exists, not whether it's large enough to matter
- Example: A difference of 0.1% in conversion rates might be statistically significant with $n = 1{,}000{,}000$ but have negligible business impact

This highlights the distinction between **statistical significance** and **practical significance**.

d. FALSE

The p-value is calculated assuming the **NULL** hypothesis is true, not the alternative. We assume H_0 , then determine how surprising our data would be under that assumption.

Question 8

A company tests whether a new packaging design affects product sales. The test produces a p-value of 0.12 at significance level $\alpha = 0.10$.

a. What decision should be made regarding the null hypothesis?

b. What type of error might have been made? Explain what this error means in context.

c. If the company really wants to detect an effect if it exists, should they increase or decrease α ? What is the trade-off?

Solution

Given:

- p-value $= 0.12$
- $\alpha = 0.10$

a. Decision:

Since p-value $(0.12) > \alpha(0.10)$, we **FAIL TO REJECT** H_0 .

b. Type of Error:

We might have made a **Type II error** (β).

Meaning: A Type II error occurs when we fail to reject a false null hypothesis.

In this context, it means:

- The new packaging design actually DOES affect sales (H_1 is true)
- But our test failed to detect this effect
- We concluded there's insufficient evidence when there really is a difference
- This could happen due to insufficient sample size, high variability, or a small effect size

c. Adjusting α :

To better detect an effect if it exists, they should **INCREASE** α (e.g., to $\alpha = 0.15$ or 0.20).

Trade-off:

- **Benefit**: Increasing α reduces the chance of Type II error (increases power) — more likely to detect a real effect
- **Cost**: Increasing α increases the chance of Type I error — more likely to incorrectly reject H_0 when it's actually true (false positive)

The trade-off is: More sensitivity to detect real effects vs. more false alarms.

Question 9

Two classification models (Model X and Model Y) are evaluated on the same test set of 500 samples:

- **Model X**: 425 correct predictions (85% accuracy)
- **Model Y**: 440 correct predictions (88% accuracy)

A statistical test comparing the models yields p-value = 0.24.

a. State the null and alternative hypotheses for comparing these models.

b. What statistical test would be most appropriate for this comparison? Why?

c. At $\alpha = 0.05$, what conclusion do you reach?

d. Model Y has higher accuracy. Why might we still not prefer it over Model X based on these results?

e. What additional information would help you make a better decision about which model to deploy?

Solution

Given:

- Model X: 425/500 correct (85%)

- Model Y: 440/500 correct (88%)
- p-value = 0.24
- Same test set

a. Hypotheses:

$$H_0 : \text{Acc}_X = \text{Acc}_Y \quad \text{(the two models have equal true accuracy)}$$

$$H_1 : \text{Acc}_X \neq \text{Acc}_Y \quad \text{(the two models have different true accuracies)}$$

Or equivalently: $H_0 : \text{Acc}_Y - \text{Acc}_X = 0$ vs. $H_1 : \text{Acc}_Y - \text{Acc}_X \neq 0$

b. Appropriate Test:

McNemar's Test is most appropriate.

Why:

- Both models are evaluated on the same test set
- We're comparing classification outcomes (correct/incorrect) on the same instances
- McNemar's test is specifically designed for paired binary outcomes
- It accounts for the dependency between the two models' predictions

Alternative: A paired test comparing the correctness vectors, or a binomial test on disagreements.

c. Conclusion at $\alpha = 0.05$:

Since p-value $(0.24) > \alpha(0.05)$, we **FAIL TO REJECT** H_0 .

Conclusion: There is insufficient statistical evidence to conclude that Model X and Model Y have different true accuracies. Although Model Y achieved 3% higher accuracy on this test set (88% vs. 85%), this difference is not statistically significant at the 0.05 level. The observed difference could reasonably be attributed to random variation in the test set.

d. Why might we not prefer Model Y?

Even though Model Y has higher observed accuracy, we might not prefer it because:

1. **Statistical uncertainty**: The 3% difference is not statistically significant, meaning it could be due to chance rather than true superiority
2. **Other considerations**:
    - Model Y might be more complex (higher computational cost, longer inference time)
    - Model Y might be harder to interpret or maintain
    - Model Y might overfit to this particular test set
    - Model X might be "good enough" for the business need
    - 3% improvement might not justify switching costs

e. Additional information needed:

1. Confidence intervals for both accuracies to understand uncertainty ranges
2. Cross-validation results or performance on multiple test sets to ensure results generalize
3. Performance on specific subgroups or error analysis (which types of errors does each model make?)
4. Computational requirements: training time, inference time, memory usage, cost
5. Model complexity: number of parameters, interpretability
6. Other metrics: precision, recall, F1-score, ROC-AUC (accuracy alone might be misleading)
7. Business impact: cost of different types of errors, real-world deployment constraints

Question 10

A regression model predicts apartment prices (in thousands of euros) based on size (in m²):

$$\hat{Y} = 45.2 + 2.8X$$

The regression output shows:

- Coefficient for X: 2.8
- Standard error: 0.35
- 95% CI for slope: [2.11, 3.49]
- p-value: < 0.001
- Sample size: $n = 75$

a. Interpret the slope coefficient in context.

b. State the null and alternative hypotheses for testing whether size has an effect on price.

c. Based on the p-value, what conclusion do you reach?

d. Interpret the 95% confidence interval for the slope.

e. A new apartment is 120 m². Predict its price and explain why you should be cautious about this prediction without additional information.

Solution

Given:

- Model: $\hat{Y} = 45.2 + 2.8X$
- Coefficient: $\beta_1 = 2.8$
- SE = 0.35
- 95% CI: [2.11, 3.49]
- p-value < 0.001
- $n = 75$

a. Interpret slope:

The slope coefficient of 2.8 means: For each additional square meter of apartment size, the predicted price increases by €2,800 (since Y is in thousands of euros).

More formally: Holding all else constant, a one-unit (1 m²) increase in apartment size is associated with an average increase of €2,800 in apartment price.

b. Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(apartment size has no linear effect on price)}$$

$$H_1 : \beta_1 \neq 0 \quad \text{(apartment size has a linear effect on price)}$$

c. Conclusion based on p-value:

Since p-value $< 0.001 < 0.05$, we **STRONGLY REJECT** H_0.

Conclusion: There is highly significant statistical evidence that apartment size has a real linear effect on apartment price. The relationship observed in our sample is extremely unlikely to have occurred by chance if there were truly no relationship in the population.

---

d. Interpret 95% CI for slope:

The 95% confidence interval [2.11, 3.49] means:

"We are 95% confident that the true increase in apartment price per additional square meter lies between €2,110 and €3,490."

Key observations:

- The entire interval is positive, confirming a positive relationship
- The interval does not contain 0, which aligns with rejecting H_0
- There is some uncertainty in the exact magnitude of the effect, but we're confident it's at least €2,110 per m²

e. Predict price for 120 m² apartment:

Prediction:

$$\hat{Y} = 45.2 + 2.8(120) = 45.2 + 336 = 381.2$$

Predicted price: €381,200

Cautions:

1. **Extrapolation risk**: We don't know if 120 m² is within the range of apartment sizes in our sample. If it's outside the range (extrapolation), the linear relationship might not hold.
2. **Prediction interval needed**: This is a point prediction. We should compute a prediction interval to understand the uncertainty. Individual predictions have much more uncertainty than the regression line itself.
3. **Model assumptions**: The prediction assumes the linear relationship holds at 120 m², no other important factors are missing, and the model's assumptions are satisfied.
4. **Simple model**: This model only uses size. Other factors (location, condition, amenities) also affect price but aren't included.

Question 11

A ride-sharing company wants to evaluate whether a new surge pricing algorithm affects driver earnings. They collect data from 60 drivers over one week:

- **Old algorithm**: $n_1 = 60$, $\bar{x}_1 = €487$, $s_1 = €95$

- **New algorithm**: $n_2 = 60$, $\bar{x}_2 = €523$, $s_2 = €102$

a. Calculate 95% confidence intervals for the mean weekly earnings under each algorithm.

b. Based solely on the confidence intervals, can you conclude that the new algorithm increases earnings? Explain your reasoning.

c. To formally test whether the new algorithm increases earnings, state the appropriate null and alternative hypotheses.

d. Explain which statistical test you would use and why.

e. Suppose the test yields p-value = 0.032. What is your conclusion at $\alpha = 0.05$?

f. The company's data scientist argues: "The p-value is significant, but the practical difference is only €36 per week. After accounting for implementation costs of €50,000, this may not be worthwhile." Discuss the distinction between statistical significance and practical significance in this context.

g. What concerns might you have about the validity of this study? Mention at least two potential issues.

Solution

Given:

- Old algorithm: $n_1 = 60$, $\bar{x}_1 = €487$, $s_1 = €95$
- New algorithm: $n_2 = 60$, $\bar{x}_2 = €523$, $s_2 = €102$
- $\alpha = 0.05$

a. 95% Confidence Intervals:

For OLD algorithm:

- $df = 60 - 1 = 59$
- $t(0.025, 59) \approx 2.001$
- $SE_1 = \frac{s_1}{\sqrt{n_1}} = \frac{95}{\sqrt{60}} = \frac{95}{7.746} \approx 12.27$
- $ME_1 = 2.001 \times 12.27 \approx 24.55$
- $CI_1 = 487 \pm 24.55$

Thus, $CI_1 = [€462.45, €511.55]$

For NEW algorithm:

- $df = 60 - 1 = 59$
- $t(0.025, 59) \approx 2.001$
- $SE_2 = \frac{s_2}{\sqrt{n_2}} = \frac{102}{\sqrt{60}} = \frac{102}{7.746} \approx 13.17$
- $ME_2 = 2.001 \times 13.17 \approx 26.35$
- $CI_2 = 523 \pm 26.35$

Thus, $CI_2 = [€496.65, €549.35]$

b. Can we conclude new algorithm increases earnings?

Based on confidence intervals alone, we **CANNOT definitively conclude** the new algorithm increases earnings, although there is suggestive evidence.

Reasoning:

- The intervals overlap (old: [462.45, 511.55], new: [496.65, 549.35])
- The overlap region is approximately [496.65, 511.55]
- However, the entire new algorithm interval is shifted higher
- Most of the new algorithm's interval is above the old algorithm's interval
- The new algorithm's lower bound (496.65) is above the old algorithm's point estimate (487)

Conclusion: While the intervals overlap, the substantial shift suggests a difference. However, overlapping confidence intervals don't provide a formal hypothesis test. We need a proper two-sample test to reach a definitive conclusion.

c. Hypotheses for formal test:

$$H_0 : \mu_{\text{new}} \leq \mu_{\text{old}} \quad \text{or} \quad \mu_{\text{new}} - \mu_{\text{old}} \leq 0$$

(New algorithm does not increase earnings)

$$H_1 : \mu_{\text{new}} > \mu_{\text{old}} \quad \text{or} \quad \mu_{\text{new}} - \mu_{\text{old}} > 0$$

(New algorithm increases earnings)

d. Which test to use and why:

Use: Two-sample t-test (Welch's t-test preferred)

Why:

- Comparing means from two independent samples
- Population standard deviations are unknown (using sample SDs)
- Need to account for potentially unequal variances ($s_1 = 95$, $s_2 = 102$)
- Welch's t-test does not assume equal variances, making it more robust
- Sample sizes are equal ($n_1 = n_2 = 60$), which is good for power

Test statistic: $t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

e. Conclusion with p-value = 0.032:

Decision: Since p-value $(0.032) < \alpha(0.05)$, we **REJECT** H_0 .

Conclusion: At the 0.05 significance level, there is statistically significant evidence that the new surge pricing algorithm increases driver weekly earnings. The observed increase of €36 (from €487 to €523) is unlikely to be due to random chance alone if the algorithms truly had equal effects.

In context: The data provide sufficient statistical evidence to support adopting the new algorithm based on its effect on driver earnings, though practical considerations (discussed in part f) should also be considered.

f. Statistical vs. Practical Significance:

Statistical Significance:

- The p-value (0.032) indicates the result is statistically significant
- This means the observed difference (€36/week) is unlikely due to random chance
- It confirms a real, detectable effect exists

Practical Significance:

- €36/week per driver = €36 × 52 = €1,872 per driver per year
- Implementation cost: €50,000
- Break-even analysis:
    - Need €50,000 / €1,872 ≈ 27 drivers to break even in one year

Key distinction:

- **Statistical significance** tells us an effect exists and is real
- **Practical significance** tells us whether the effect is large enough to matter for decision-making
- With large samples, even tiny effects become statistically significant

- A result can be statistically significant but practically insignificant (or vice versa)

Other considerations:

- Long-term effects: Will the €36/week persist over time?
- Driver satisfaction and retention
- Competitive advantages
- Maintenance and ongoing costs
- Customer impact of surge pricing

Bottom line: Statistical significance is necessary but not sufficient for making business decisions. We must also consider effect size, costs, and broader impacts.

g. Validity concerns:

Potential issues with this study:

1. **Selection bias / Non-random assignment**:
    - Were drivers randomly assigned to algorithms?
    - Did drivers self-select into groups?
    - Are the drivers in each group comparable in experience, location, etc.?
    - If not randomized, pre-existing differences could confound results
2. **Temporal confounding**:
    - Were both algorithms tested during the same time period?
    - If tested sequentially, seasonal effects, economic changes, or events could explain differences
    - One week might not be representative (holidays, special events, weather)
3. **Sample representativeness**:
    - Are these 60 drivers representative of all drivers?
    - Different cities, demographics, experience levels might respond differently
    - Results might not generalize to the broader driver population
4. **Learning effects / Novelty**:
    - Drivers might need time to adapt to the new algorithm
    - Initial results might not reflect long-term performance
    - Hawthorne effect: drivers might behave differently knowing they're being studied
5. **External validity**:
    - One-week observation period might be too short
    - Doesn't capture week-to-week or seasonal variability
    - Long-term driver behavior and earnings patterns unclear

---

6. **Independence assumption**:
    - Are the 60 drivers truly independent?
    - Drivers in the same city might be affected by same market conditions
    - Violates independence assumption if there's clustering

Recommendations to address concerns:

- Conduct a randomized controlled trial (RCT)
- Extend observation period to multiple weeks
- Include diverse driver populations across locations
- Monitor for confounding variables
- Consider a crossover design where drivers experience both algorithms

Bonus Question 1

Explain why we use $n-1$ instead of n in the denominator when calculating sample variance s^2. What property does this give our estimator?

Solution

Why $n-1$ instead of n?

When calculating sample variance, we use:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{instead of} \quad \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

Reasons:

1. **Degrees of freedom**: We "use up" one degree of freedom when we estimate the mean ($\bar{x}$) from the same data. Once we know $(n-1)$ deviations and the mean, the n th deviation is determined. We have only $(n-1)$ independent pieces of information.
2. **Bias correction**: Using n in the denominator produces a biased estimator that systematically underestimates σ^2. This is because:
    - We're measuring deviations from $\bar{x}$ (sample mean), not μ (population mean)
    - $\bar{x}$ is the value that minimizes $\sum(x_i - \bar{x})^2$
    - Deviations from $\bar{x}$ are systematically smaller than deviations from μ

- Using $(n-1)$ compensates for this underestimation

3. **Unbiasedness property**: Using $(n-1)$ makes s^2 an **unbiased estimator** of σ^2, meaning:

$$E[s^2] = \sigma^2$$

- On average, across many samples, s^2 equals the true population variance
- This is called "Bessel's correction"

Mathematical insight:

$$E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = (n-1)\sigma^2$$

Therefore:

$$E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right] = \sigma^2$$

In practice: The difference between n and $(n-1)$ is negligible for large samples but important for small samples.

Bonus Question 2

A researcher conducts 20 different hypothesis tests, each at $\alpha = 0.05$. If all null hypotheses are actually true, approximately how many would you expect to be rejected just by chance? What problem does this illustrate, and what is one approach to address it?

Solution

Expected false positives:

With $\alpha = 0.05$ and 20 tests where all H_0 are true:

$$\text{Expected rejections} = 20 \times 0.05 = 1 \text{ test}$$

We'd expect about **1 false positive** (Type I error) just by chance.

Problem illustrated: **Multiple Testing Problem** (Family-Wise Error Rate)

When conducting many hypothesis tests:

- The probability of at least one false positive increases with the number of tests
- $P(\text{at least one false positive}) = 1 - (1 - \alpha)^k$ where k = number of tests

For $k = 20$, $\alpha = 0.05$:

$$P = 1 - (0.95)^{20} = 1 - 0.358 = 0.642 \text{ or } 64.2\%$$

Nearly 2/3 chance of at least one false discovery!

Approaches to address this:

1. **Bonferroni Correction**:
    - Divide α by number of tests
    - Use $\alpha^* = 0.05/20 = 0.0025$ for each test
    - Very conservative; reduces power significantly
    - Good when tests are independent
2. **False Discovery Rate (FDR) Control**:
    - Benjamini-Hochberg procedure
    - Controls proportion of false discoveries among rejections
    - Less conservative than Bonferroni
    - Better for exploratory research
3. **Holm-Bonferroni Method**:
    - Sequential rejection procedure
    - More powerful than standard Bonferroni
    - Orders p-values and uses adjusted thresholds
4. **Reduce number of tests**:

- Only test pre-specified hypotheses
- Avoid data dredging / p-hacking
- Use theory to guide hypothesis selection

Practical implications:

- Always report how many tests were conducted
- Pre-register hypotheses when possible
- Be transparent about multiple testing
- Consider adjustment methods appropriate for your context
- Distinguish confirmatory from exploratory analyses

Quick Reference Formulas

Concept	Formula
CI for mean (σ unknown)	$\bar{x} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$
CI for mean (σ known)	$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
CI for proportion	$\hat{p} \pm z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Test statistic for mean	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
Test statistic for proportion	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Standard error (mean)	$SE = \frac{s}{\sqrt{n}}$
Standard error (proportion)	$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Sample variance	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$