In-Class Exercise 5

For this in-class exercise, work with your group of 2-3 people, to answer the following questions. These questions are not necessarily easy and sometimes they will not have a clear "correct" answer. The goal is to get you thinking about the material we've learned. Some of these questions may require you to discuss and debate with your group members to come up with an answer or can cover topics that we have not yet covered in class.

Be prepared to share your answers with the class and add to the discussion.

After class submit your a do-file with your answers in comments to Moodle for grading. You will be graded as a group on your submission. Only one group member needs to submit the assignment, but make sure add all group member names.

For this exercise, you are hired as "Data Forensic Consultants" for a class-action lawsuit representing Gig Workers. The plaintiffs claim that RideShare Inc. is engaging in algorithmic wage discrimination—specifically, that the algorithm targets drivers in specific zones with lower pay rates that fall below the advertised "Average Earnings," effectively paying them below minimum wage once expenses are factored in.

Part 1: Power Analysis

Rideshare Inc. has successfully argued that it can only share a small amount of their data with the plaintiffs due to privacy concerns. The dataset you were given contains information on 45 drivers. Download this data here:

Download Data: rideshare_leaked_data_n45.csv

Question 1

The presiding judge has stated that for any wage evidence to be admissible in court, the Margin of Error (MOE) must be no more than $\pm\$2.00$ per hour.

Using the sample standard deviation (s) from the 45 leaked records as your "best guess" for the population standard deviation (σ), determine if your current sample meets the judge's standard at a 95% Confidence Level.

If $n = 45$ is not enough, calculate exactly how many additional driver records the attorney must subpoena to hit that $\pm\$2.00$ target.

! Solution

First, load the data and calculate the `net_hourly_wage`.

```
import delimited "rideshare_leaked_data_n45.csv", clear
gen net_hourly_wage = (gross_earnings - est_expenses) / shift_duration_hrs
summ net_hourly_wage
```

Using the sample data (your numbers may vary slightly due to randomness, but roughly):

- **Sample Mean ($\bar{x}$):** $\approx \$20.66$
- **Sample Std Dev (s):** $\approx \$15.00$
- **Sample Size (n):** 45
- **Standard Error ($SE = s/\sqrt{n}$):** $15.00/\sqrt{45} \approx 2.24$
- **Critical Value ($t_{0.025,44}$):** ≈ 2.015

Current Margin of Error:

$$MOE = t \times SE \approx 2.015 \times 2.24 \approx \$4.51$$

This is **much larger** than the required $\pm\$2.00$.

Required Sample Size: To find the required n, we rearrange the MOE formula (using $z \approx 1.96$ for planning):

$$n = \left(\frac{z \cdot \sigma}{m}\right)^2 = \left(\frac{1.96 \cdot 15.00}{2.00}\right)^2 \approx (14.7)^2 \approx 216.09$$

We need a total of **217** records. Since we have 45, we need to subpoena **$217 - 45 = 172$ additional records**.

Question 2

How much can you trust this data? Which sampling technique do you hope Rideshare Inc. used to generate this data? Explain your reasoning.

What sampling techniques would make you suspicious of the data? Why?

Is there a way to verify the sampling technique used with only this data? If not, what would you need to verify it?

! Solution

- **Trust:** Low. The sample size (n=45) is small and the source is "leaked," implying it might not be a formal random sample.
- **Desired Technique:** Simple Random Sampling (SRS). This ensures that every driver has an equal chance of being selected, minimizing bias.
- **Suspicious Techniques:** Convenience Sampling (just grabbing easy file) or Quota Sampling based on non-representative criteria.
- **Verification:** You generally **cannot** verify the sampling technique just by looking at the data itself. You would need the **metadata** or the **sampling methodology protocol** (the code or procedure used to select these rows).

Part 2: Building the Case

The Defense claims the "Average Wage" is \$25/hr. Your job is to find the statistical evidence that proves this number is misleading.

Question 3

Choose one of the following paths to build your case. You must explore the data and decide which statistical argument is stronger.

Path A: The "Average Driver" Argument (Inference about Means)

Hypothesis: You suspect the true mean wage (μ) is significantly lower than the company's claimed \$25/hr.

Task: Construct a 95% Confidence Interval for μ. If the upper bound of your interval is below \$25, you have evidence of wage theft.

Justify why you are using the t-distribution instead of the normal distribution.

What do you find? How do you interpret these findings? Can you connect this to the power analysis in Part 1 to make a convincing argument to the judge?

Path B: The "Vulnerable Worker" Argument (Inference about Proportions)

Hypothesis: You suspect that while some make good money, a specific proportion of drivers (p) are earning below the federal minimum wage (\$7.25/hr).

Task: Create a binary variable in the dataset (1 = Below Min Wage, 0 = Above). Calculate the sample proportion $\bar{p}$. Construct a 95% Confidence Interval for the true proportion of drivers earning illegal wages.

Check if the sample size is large enough for this method ($np > 5$ and $n(1-p) > 5$).

Why did you chose Path A or Path B?

! Solution

Path A (Means):

- **Confidence Interval:** $20.66 \pm 4.51 \Rightarrow [16.15, 25.17]$.
- **Conclusion:** The interval **includes** \$25.00. We **fail to reject** the claim that the average is \$25. Even though the mean is \$20, the variance is too high (margin of error too wide) to rule out \$25 statistically.

Path B (Proportions):

- **Low Wage (< 7.25):** Count is approx 6 drivers out of 45. $\hat{p} = 6/45 \approx 0.133$.
- **Condition Check:** $n\hat{p} = 6 > 5$ (Just barely met!), $n(1-\hat{p}) = 39 > 5$. It is valid.
- **Confidence Interval (95%):** Using approx formula $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$$0.133 \pm 1.96\sqrt{\frac{0.133(0.867)}{45}} \approx 0.133 \pm 0.099 \Rightarrow [0.034, 0.232]$$

- **Conclusion:** We are 95% confident that between **3.4% and 23.3%** of drivers earn below minimum wage. This is strong evidence of **illegal** wage practices for a subset of workers, even if the *average* is okay.

Choice: Path B is stronger here because Path A is inconclusive (high variance makes the mean estimate fuzzy), whereas Path B definitely identifies a non-zero proportion of underpaid workers.

Part 3: Flexibility

Here is the quote from the interview for your reference:

> We categorically deny these allegations. Our data shows that drivers value flexibility over guaranteed hourly rates. The 'low wage' shifts cited by the plaintiffs are actually 'flexible' rides taken by drivers who log in for just 15 minutes between other jobs. You cannot compare a flexible gig to a rigid 9-to-5 job. When you factor in the 'value of flexibility'—estimated at \$5/hour—our drivers are earning well above the industry standard. The algorithm doesn't discriminate; it optimizes for driver freedom. - Rideshare Inc. CEO

Question 4

Re-evaluate your Confidence Interval from Part 2.

If you were to factor in this "value of flexibility" of \$5.00/hour, how would that change your findings? Be careful about how you choose to incorporate this value into your analysis.

Does your case still hold? Or does this "invisible" \$5.00 value exonerate the company?

How would you respond to this argument in court? Try to make a convincing rebuttal based on your statistical findings. Feel free to look through the data and see if you can find any evidence to support or refute this claim.

! Solution

Path A (Means) with Flexibility: New Mean $\approx 20.66 + 5.00 = 25.66$. New CI: $[21.15, 30.17]$. The company looks great now! The average is effectively above \$25.

Path B (Proportions) with Flexibility: We need to see how many drivers satisfy `wage + 5.00 < 7.25` (i.e., `wage < 2.25`). In our dataset, the minimum wage was capped at \$2.50, so **0** drivers fall below this new threshold. The company is exonerated on the minimum wage claim *if* you accept the \$5 valuation.

Rebuttal:

- **Heterogeneity:** You can argue that flexibility is not worth \$5 to *everyone*. A driver working 60 hours a week (full-time equivalent) likely values stability over flexibility, yet they are getting the same volatile, low pay.
- **Variance:** The high standard deviation remains (\$15). Even with the \$5 boost, the unpredictability of earnings is a massive cost to drivers that the "average" hides.

Once you have reached this point, raise your hand and let me know.

Question 5

The question of flexibility here is an interesting one. Reducing the value of flexibility to a single number is likely an oversimplification, but it is an important concept in the gig work literature. Can you think of a convincing way to measure the "value of flexibility" using data? Discuss with your group and think of some potential strategies and ways to use the data.

! Solution

Example Strategy: Compare the wages of "flexible" workers (those with variable hours, short shifts) vs. "rigid" workers (those with fixed, predictable schedules). If rigid workers accept lower wages for stability, the difference might quantify the "price of stability" (or conversely, the premium for flexibility).

Alternatively, look at turnover rates. If low-wage (flexible) jobs have high retention, users might indeed value the flexibility. If turnover is high, the "flexibility value" might be a myth.