
Midterm Project

Applied Statistical Inference

Purpose

Use data to answer questions about airport satisfaction and coffee-shop spending. This project assesses your ability to clean data, select a method, verify assumptions, and explain results in language that a decision-maker can use.

Part A: SFO Passenger Satisfaction

Use `SFO 2009.csv`. Ratings in columns E through K use a 1 to 5 scale, where 1 is Poor and 5 is Excellent. Values greater than 6 represent missing data. Convert those values to `NA` before analysis and report how many values were changed in each variable you use.

1. Create labeled boxplots for airport, restaurant, retail, and safety ratings. Compare the distributions without treating the ordinal ratings as more precise than they are.
2. Define p_S as the population proportion of passengers who rate safety at least 4. Estimate p_S and compute a 95% confidence interval.
3. Define p_A as the population proportion of passengers who rate the airport at least 4. Estimate p_A and compute a 95% confidence interval.
4. Explain why overlap or non-overlap of two separate confidence intervals is not, by itself, the correct test of $p_S = p_A$.
5. Because the same passenger rates both items, use a method that respects the paired binary responses to compare the two proportions. State the hypotheses, report the result, and interpret it.

Part B: Coffee-Shop Spending

Use `Coffee Shop Spending.csv`, which contains spending at coffee shops A and B.

1. Inspect the data structure and determine whether the samples are independent or paired. Support your decision from the dataset.
2. Estimate the population mean spending for each shop and report 95% confidence intervals.
3. Test whether population mean spending differs between the shops using the design identified above. Report the estimated difference and its 95% confidence interval.
4. Define a high-spending visit as more than \$12. Estimate the high-spending proportion for each shop.
5. Compare the proportions using a method appropriate to the study design. Report and interpret a 95% confidence interval for the difference.

6. Write a short management summary. State what the evidence suggests and identify one limitation of the analysis.

Submission Checklist

Submit through WebCampus (Canvas) by the deadline posted there. Submit both of the following files:

- a rendered PDF report; and
- the complete `.Rmd` source file used to produce the report.

The PDF must include your name and assignment title, answer every question in the order presented, show readable and labeled tables and figures, and explain results in complete sentences. The `.Rmd` file must contain all code needed to reproduce the report. Both files must run from the same data supplied for the assignment. Do not submit unexplained software output. You are responsible for checking all results and explaining every choice you made.

A Note About AI

AI can be used as a resource, but it cannot do the assignment for you. I do not want to read AI slop. Do not paste pages of generated writing, code, or interpretations into your report and expect me to sort through it. In the real world, you cannot send AI garbage to a boss, client, or customer. You should not submit it in this class either.

If you use AI, use it to help you learn. Check what it gives you. Make sure the code works. Internalize the material, interpret it, and then explain it in your own words and voice. You should understand every part of your submission and be able to defend the choices you made. Include a short note at the end of the report that names the tool and tells me how you used it. If the work reads like unreviewed AI output, it will receive a 0%, and you will have the opportunity to redo it under the policy in the syllabus.

How Your Work Will Be Evaluated

Credit is based on the statistical reasoning, correct use of R, appropriate checks of assumptions, quality of visual and numerical evidence, and accuracy of the interpretation. A correct number without an explanation is not a complete answer.