BZAN 535: Homework 2

Joseph Falcone

2026-09-20

Statistical Inference for Household Shopping Behavior

Use `transaction_data.csv` from Homework 1 and `hh_demographic.csv`. For each question, decide what quantity answers the business question and which inferential method is appropriate. Support your conclusion with an estimate, a confidence interval, or a hypothesis test as requested.

About the Data

- The transaction fields have the same meanings as in Homework 1. A unique `BASKET_ID` represents one observed shopping trip.
- `household_key` identifies households in both files.
- `WEEK_NO` identifies the week within the observation period.
- `hh_demographic.csv` contains demographic information for a subset of the households. `INCOME_DESC` identifies the reported income category.

Submission and Expectations

- Submit your **.rmd source and knitted PDF**, with your name on the report. Each question is worth **5 points**, for a total of **20 points**.
- Include your analysis code and the summaries, visualizations, and inferential results needed to support your conclusions. Avoid repeating exploratory work that does not contribute to the question.
- State the population quantity you are estimating and justify your choice of method. For comparisons, make the direction of subtraction clear.
- Use 95% confidence intervals and a significance level of $\alpha = 0.05$. For each hypothesis test, state the null and alternative hypotheses.
- Examine whether the assumptions of your method are reasonable. Explain important limitations, including which households or settings your conclusions might reasonably apply to.
- Interpret results in business terms. Distinguish evidence of a difference from its practical importance, and explain what the analysis does and does not establish.
- Different approaches can receive full credit when the analysis is correct, the choices are defensible, and the conclusions follow from the evidence.

- Before consulting an AI tool, sketch your own approach to each question. You may then use AI to support your work under the course policy. You are responsible for verifying the analysis and explaining the choices and conclusions in your submission.

1. How Much Does a Household Spend? (5 points)

Leadership wants an estimate of mean total household spending at Kroger over the observation period represented in the transaction file.

Provide a point estimate and an appropriate 95% confidence interval. Explain what the interval says about average household spending and what it does not say about the spending of an individual household.

Discuss one important limitation of using these observed households and their recorded purchases to make a broader population statement.

```
household_spending <- transactions |>
    group_by(household_key) |>
    summarise(
        total_spend = sum(SALES_VALUE, na.rm = TRUE),
        .groups = 'drop'
    )

t.test(household_spending$total_spend, conf.level = 0.95)
```

```
    One Sample t-test

data:  household_spending$total_spend
t = 48.118, df = 2499, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 3091.642 3354.328
sample estimates:
mean of x
 3222.985
```

Response: The estimated mean total household spending during the observational period is 3222.99, with a 95% confidence interval of (3091.642, 3354.328). This interval suggests that we are 95% confident that the true population mean household spending is within this range. The sample mean of 3222.99 serves as our point estimate of the population mean, while the confidence interval reflects the uncertainty associated with that estimate. This interval does not suggest that 95% of individual households spent between 3091.642 and 3354.328, but rather reflects the population collectively. One limitation worth mentioning is the potential for sampling bias or lack of representativeness.

Because we do not know whether the households included in the transaction data are representative of the broader population of Kroger households, the observed spending patterns may differ from those of customers who were not included in the sample. Therefore, the estimated population mean may not generalize perfectly to all Kroger households.

---

2. Has Spending per Trip Changed? (5 points)

Leadership wants to know whether household spending per trip changed from the first year of data to the period that follows (weeks 1–52 versus weeks 53–102). Compare weeks 1–52 with weeks 53–102. Set up the comparison yourself: decide what the unit of analysis should be, which households or trips to include, and whether the two periods are better treated as independent groups or as linked observations. Justify that decision using what you know about how these data were generated. Conduct an appropriate hypothesis test and provide a 95% confidence interval for the difference. Explain how your setup determines your choice of method and what the result suggests about shopping behavior. Would the result establish that a change in Kroger's business practices caused spending per trip to change? Explain.

```
basket_spending <- transactions |>
    filter(WEEK_NO >= 1, WEEK_NO <= 102) |>
    mutate(
        period = if_else(WEEK_NO <= 52, 'first', 'second')
    ) |>
        group_by(household_key, BASKET_ID, period) |>
        summarise(trip_spend = sum(SALES_VALUE, na.rm = TRUE),
        .groups = 'drop'
        )

household_period <- basket_spending |>
    group_by(household_key, period) |>
    summarise(
        mean_trip_spend = mean(trip_spend),
        .groups = 'drop'
    )

paired_households <- household_period |>
    pivot_wider(
        names_from = period,
        values_from = mean_trip_spend
    ) |>
        drop_na(first, second)

paired_households <- paired_households |>
    mutate(
```

```
        difference = second - first
    )

paired_households |>
    summarise(
        mean_first = mean(first),
        mean_second = mean(second),
        mean_difference = mean(difference)
    )
```

```
# A tibble: 1 x 3
  mean_first mean_second mean_difference
       <dbl>       <dbl>           <dbl>
1       31.2        32.2            1.07
```

```
t.test(
    paired_households$second,
    paired_households$first,
    paired = TRUE,
    conf.level = 0.95
)
```

```

    Paired t-test

data:  paired_households$second and paired_households$first
t = 3.6168, df = 2475, p-value = 0.0003042
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 0.4904119 1.6519166
sample estimates:
mean difference
       1.071164
```

Response: Households observed in both periods are treated as linked observations because the same households can make multiple shpping trips and appear in both the first year (weeks 1-52) and second year (weeks 53-102). For each household, average spending per trip in each time period (year one and year two) was calculated and then compared using a paired t-test. This resulted in 2476 households with observations in both periods. Mean spending was approximately 31.20 in weeks 1-52 and 32.20 in weeks 53-102. The estimated mean increase was 1.07 per trip, with a 95% confidence interval of (0.49, 1.65). The paired t-test produced a p-value of 0.000304, providing evidence that the mean spending per trip differed between the two periods. Because the entire confidence interval is above zero (i.e. zero is outside of the interval), the data suggests that households spent, on average, more per trip during the second period.

---

3. Does Spending Differ by Income Group? (5 points)

Marketing wants to understand whether average total household spending over the observation period differs between two income groups. Choose two categories from `INCOME_DESC` and explain why the comparison is useful. Make this choice before examining spending differences between the categories. Category sizes vary considerably,so check how many households fall in each before committing, and comment on how those group sizes affect the precision of your estimate. Conduct an appropriate hypothesis test and provide a 95% confidence interval for the population mean difference. Interpret both the statistical evidence and the size of the estimated difference. What can marketing reasonably learn from this comparison, and what would remain uncertain about a promotion targeted by income?

```
income_spending <- household_spending |>
    inner_join(demographics, by = 'household_key')

income_spending |>
    count(INCOME_DESC, sort = TRUE)
```

```
# A tibble: 12 x 2
   INCOME_DESC     n
   <chr>       <int>
 1 50-74K        192
 2 35-49K        172
 3 75-99K         96
 4 25-34K         77
 5 15-24K         74
 6 Under 15K      61
 7 125-149K       38
 8 100-124K       34
 9 150-174K       30
10 175-199K       11
11 250K+          11
12 200-249K        5
```

```
two_income_groups <- income_spending |>
    filter(INCOME_DESC %in% c('50-74K', '100-124K')) |>
    mutate(
        INCOME_DESC = factor(
            INCOME_DESC,
            levels = c('100-124K', '50-74K')
        )
    )

two_income_groups |>
```

```
    group_by(INCOME_DESC) |>
    summarise(
        n = n(),
        mean_spend = mean(total_spend),
        sd_spend = sd(total_spend),
        .groups = "drop"
    )
```

```
# A tibble: 2 x 4
  INCOME_DESC     n mean_spend sd_spend
  <fct>       <int>      <dbl>    <dbl>
1 100-124K       34      5928.    3501.
2 50-74K        192      5702.    3284.
```

```
t.test(
    total_spend ~ INCOME_DESC,
    data = two_income_groups,
    conf.level = 0.95
)
```

```
    Welch Two Sample t-test

data:  total_spend by INCOME_DESC
t = 0.34918, df = 43.906, p-value = 0.7286
alternative hypothesis: true difference in means between group 100-124K and group 50-
74K is not equal to 0
95 percent confidence interval:
 -1075.553  1526.325
sample estimates:
mean in group 100-124K   mean in group 50-74K
              5927.740               5702.355
```

Response: `50-74K` and `100-124K` income categories were used because they provide a meaningful comparision between moderate and higher income households while still retaining usable sample sizes. The `50-74K` group, refered to hereafter as 'group 1' contains 192 households and the `100-124K` group, refered to hereafter as 'group 2' contains 34. Because the higher-income group is much smaller, its population mean is estimated with less precision, which contributes to a relatively wide confidence interval for the difference. Average total household spending was 5,702.36 for group 1 and $5,927.74 for group 2. The estimated difference, defined as group 2 minus group 1, was approximately 225.39. A Welch two-sample t-test produced a p-value of 0.7286, with a 95% confidence interval for the population mean difference of (-$1,075.55, $1,526.33). Because this interval includes zero and the p-value is large, there is not sufficient statistical evidence that mean total

spending differs between these two income groups. The estimated difference itself is also small relative to average spending in either group, although the wide confidence interval means the true difference could possibly be meaningfully negative or positive. Marketing can reasonably conclude that these data do not show a clear spending difference between these two income categories. However, this comparison does not establish whether an income targeted promotion would be effective. Data is observational and doesn't measure customers' actual responses to a promotion.

---

4. Can the Demographic Subset Speak for Everyone? (5 points)

Marketing often bases its analyses on households with demographic information available. Leadership wants to know whether these households resemble those without demographic information in their shopping behavior. Choose one relevant household shopping measure and explain your choice. Compare households with and without a record in `hh_demographic.csv`, using an appropriate hypothesis test and a 95% confidence interval. Based on your findings, how comfortable would you be extending results from the demographic subset to all households in the transaction data? Would finding no statistically significant difference on your chosen measure establish that the subset is representative? Explain.

```
household_demo_status <- household_spending |>
    mutate(
        demo_available = if_else(
            household_key %in% demographics$household_key,
            'record found',
            'no record found'
        )
    )

household_demo_status |>
    group_by(demo_available) |>
    summarise(
        n = n(),
        mean_spend = mean(total_spend),
        sd_spend = sd(total_spend),
        .groups = 'drop'
    )
```

```
# A tibble: 2 x 4
  demo_available      n mean_spend sd_spend
  <chr>           <int>      <dbl>    <dbl>
1 no record found  1699      2095.    2430.
2 record found      801      5615.    3751.
```

```r
t.test(
    total_spend ~ demo_available,
    data = household_demo_status,
    conf.level = 0.95
)
```

```
    Welch Two Sample t-test

data:  total_spend by demo_available
t = -24.266, df = 1127.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group no record found and group record
95 percent confidence interval:
 -3804.535 -3235.315
sample estimates:
mean in group no record found    mean in group record found
                      2095.201                        5615.126
```

Response: Total household spending over the observation period was chosen as the shopping measure because it directly reflects the overall value of a household's purchasing activity and is relevant to many marketing analyses. Households were divided into those with a record in hh_demographic.csv and those without one. The demographic-record group contained 801 households, while the group without demographic information contained 1,699 households. Average total household spending was $5,615.13 among households with demographic information and $2,095.20 among households without demographic information. A Welch two-sample t-test found a statistically significant difference between the groups, with a p-value of less than 2.2 × 10^-16. The estimated difference was approximately $3,519.93 higher for households with demographic records, with a 95% confidence interval of approximately ($3,235.32, $3,804.54). These results suggest that households with demographic information differ substantially from households without demographic information in total observed spending. Therefore, I would be cautious about extending results from the demographic subset to all households in the transaction data, since the subset appears to contain considerably heavier-spending households. Even if no statistically significant difference had been found for total spending, that alone would not establish that the demographic subset is representative. Representativeness involves more than one characteristic, and the two groups could still differ in shopping frequency, spending per trip, product mix, coupon usage, or other observed and unobserved characteristics.
