

Homework 4

Problem 1: NBC pilot survey

Like any TV network, NBC conducts market research on how viewers respond to TV shows (both its own shows and the shows of other competing networks). The data in `nbc_pilotsurvey.csv` contains the results of some of that research. Each row of this data frame shows the responses of a single viewer (the `Viewer` variable) to the “pilot” episode¹ of a single TV show (`Show` variable). The remaining variables encode the viewers reactions to the show. Viewers were asked to rate the strength of their agreement on a 1-5 scale (where 5 means “strongly agree”) with various statements about the show, such as “This show made me feel happy” or “I found this show confusing.”²

Use this data to answer the questions below. For each Part (A, B, C), your response should include four sections:

- 1) Question: What question are you trying to answer?
- 2) Approach: What approach/statistical tool did you use to answer the question?
- 3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.)
- 4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

These questions are fairly simple, so we’d expect each of these four sections for each part to be quite short—surely no more than 1-3 sentences each.

Your confidence intervals can be constructed either using bootstrap or using a built-in R function that reports “large sample” confidence intervals (i.e. based on the Central Limit Theorem). **While you shouldn’t include raw R code, make sure to state which approach you used in each “Approach” section.**

Part A. Consider the shows “Living with Ed” and “My Name is Earl.” Who makes people happier: Ed or Earl? Construct a filtered data set containing only viewer responses where `Show == "Living with Ed"` or `Show == "My Name is Earl"`. Then construct a 95% confidence interval for the difference in mean viewer response to the `Q1_Happy` question for these two shows. Is there evidence that one show consistently produces a higher mean `Q1_Happy` response among viewers?

Part B. Consider the shows “The Biggest Loser” and “The Apprentice: Los Angeles.” Which reality/contest show made people feel more annoyed? Construct a filtered data set containing only viewer responses where `Show == "The Biggest Loser"` or `Show == "The Apprentice: Los Angeles"`. Then construct a 95% confidence interval for the difference in mean viewer response to the `Q1_Annoyed` question for these two shows. Is there evidence that one show consistently produces a higher mean `Q1_Annoyed` response among viewers?

Part C. Consider the show “Dancing with the Stars.” This show has a straightforward premise: it is a dancing competition between couples, with each couple consisting of a celebrity paired with a professional dancer. Per Wikipedia: “Each couple performs predetermined dances and competes against the others for judges’ points and audience votes.”

Despite the simplicity of this format, it seems that some Americans nonetheless find the show befuddling, as evidenced by our survey data on the `Q2_Confusing` question, which asked survey respondents to agree or disagree with the statement “I found this show confusing.” Any response of 4 or 5 indicated that the survey participant either Agreed (4) or Strongly Agreed (5) that “Dancing with the Stars” was a confusing show.

Construct a filtered data set containing only viewer responses where `Show == "Dancing with the Stars"`. Assuming this sample of respondents is representative of TV viewers more broadly, what proportion of

¹I.e. the first episode of the show ever made.

²In fact, all the questions labeled Q1 were “This show made me feel...” questions, whereas all the questions labeled Q2 were “I found this show...” questions.

American TV watchers would we expect to give a response of 4 or greater to the “Q2_Confusing” question? Form a 95% confidence interval for this proportion and report your results.

Problem 2: EBay

In this problem, you’ll analyze data from an experiment run by EBay in order to assess whether the company’s paid advertising on Google’s search platform was improving EBay’s revenue. (It was certainly improving Google’s revenue!) In fiscal year 2020, more than 80% of Google’s reported \$182 billion in revenue came from its advertising system. Google AdWords has advertisers bid on certain keywords (e.g., “iPhone” or “toddler shoes”) in order for their clickable ads to appear at the top of the page in Google’s search results. These links are marked as an “ad” by Google, and they’re distinct from the so-called “organic” search results that appear lower down the page.

Nobody pays for the organic search results; pages get featured here if Google’s algorithms determine that they’re among the most relevant pages for a given search query. But if a customer clicks on one of the sponsored “Ad” search results, Google makes money. Suppose, for example, that EBay bids \$0.10 on the term “vintage dining table” and wins the bid for that term. If a Google user searches for “vintage dining table” and ends up clicking on the sponsored EBay link from the page of search results, EBay pays Google \$0.10 (the amount of their bid).³

For a small company, there’s often little choice but to bid on relevant Google search terms; otherwise their search results would be buried. But a big site like EBay doesn’t necessarily have to pay in order for their search results to show up prominently on Google. They always have the option of “going organic,” i.e. **not** bidding on any search terms and hoping that their links nonetheless are shown high enough up in the organic search results to garner a lot of clicks from Google users. So the question for a business like EBay is, roughly, the following: does the extra traffic brought to our site from paid search results—above and beyond what we’d see if we “went organic”—justify the cost of the ads themselves?

To try to answer this question, EBay ran an experiment in May of 2013. For one month, they turned off paid search in a random subset of 70 of the 210 designated market areas (DMAs) in the United States. A designated market area, according to Wikipedia, is “a region where the population can receive the same or similar television and radio station offerings, and may also include other types of media including newspapers and Internet content.” Google allows advertisers to bid on search terms at the DMA level, and it infers the DMA of a visitor on the basis of that visitor’s [browser cookies](#) and IP address. Examples of DMAs include “New York,” “Miami-Ft. Lauderdale,” and “Beaumont-Port Arthur.” In the experiment, EBay randomly assigned each of the 210 DMAs to one of two groups:

- the *treatment* group, where *advertising on Google AdWords for the whole DMA was paused for a month, starting on May 22.*
- the *control* group, where *advertising on Google AdWords continued as before.*

In `ebay.csv` you have the results of the experiment. The columns in this data set are:

- `DMA`: the name of the designated market area, e.g. New York
- `rank`: the rank of that DMA by population
- `tv_homes`: the number of homes in that DMA with a television, as measured by the market research firm Nielsen (who defined the DMAs in the first place)
- `adwords_pause`: a 0/1 indicator, where 1 means that DMA was in the treatment group, and 0 means that DMA was in the control group.
- `rev_before`: EBay’s revenue in dollars from that DMA in the 30 days before May 22, before the experiment started.
- `rev_after`: EBay’s revenue in dollars from that DMA in the 30 days beginning on May 22, after the experiment started.

³There’s huge variability in the market price of different search terms. The market price per click for a search term like “insurance” or “attorney” or “MBA programs” might be \$50 or more. Google makes a *fortune* on these popular search terms. For stuff you might buy on EBay, the market price is usually a lot less.

The outcome variable of interest is the **revenue ratio** at the DMA level, i.e. the ratio of revenue after to revenue before for each DMA. If EBay's paid search advertising on Google was driving extra revenue, we would expect this revenue ratio to be systematically lower in the treatment-group DMAs versus the control-group DMAs. On the other hand, if paid search advertising were a waste of money, then we'd expect the revenue ratio to be basically equal in the control and treatment groups. Two explanatory notes here:

- We use the ratio rather than the absolute difference because the DMAs differ enormously in population and therefore revenue.
- We wouldn't necessarily expect the before-and-after revenue ratio to be 1 (i.e. similar revenue before and after the experiment), even in the control-group DMAs. That's because, like any retailer, EBay's sales exhibit a lot of seasonal patterns and might be lower in some months across the board, regardless of paid search. That's why the important question isn't whether the revenue is the same before and after in the treatment-group DMAs, but whether the before-and-after **ratio** is the same for the treatment group as for the control group.

Your task is compute the difference in revenue ratio between the treatment and control DMAs and provide a 95% confidence interval (by any means we've learned) for the difference. Use these results to assess the evidence for whether the revenue ratio is the same in the treatment and control groups, or whether instead the data favors the idea that paid search advertising on Google creates extra revenue for EBay. Make sure you use at least 10,000 Monte Carlo simulations in any bootstrap simulations.

Your write-up for this problem should include four sections:

- 1) Question: What question are you trying to answer?
- 2) Approach: What approach/statistical tool did you use to answer the question?
- 3) Results: What evidence/results did your approach provide to answer the question? (E.g. any numbers, tables, figures as appropriate.)
- 4) Conclusion: What is your conclusion about your question? Provide a written interpretation of your results, understandable to stakeholders who might plausibly take an interest in this data set.

It is certainly possibly in this case for each of these four sections to be only 1-3 sentences long, although you can take longer if you feel you need it.

Problem 3 - Iron Bank

The Securities and Exchange Commission (SEC) is investigating the Iron Bank, where a cluster of employees have recently been identified in various suspicious patterns of securities trading that violate federal "insider trading" laws. Here are few basic facts about the situation:

- Of the last 2021 trades by Iron Bank employees, 70 were flagged by the SEC's detection algorithm.
- But trades can be flagged every now and again even when no illegal market activity has taken place. In fact, the SEC estimates that the baseline probability that any *legal* trade will be flagged by their algorithm is 2.4%.
- For that reason, the SEC often monitors individual and institutional trading but does not investigate incidents that look plausibly consistent with random variability in trading patterns. In other words, they won't investigate unless it seems clear that a cluster of trades is being flagged at a rate significantly higher than the baseline rate of 2.4%.

Are the observed data (70 flagged trades out of 2021) consistent with the SEC's null hypothesis that, over the long run, securities trades from the Iron Bank are flagged at the same 2.4% baseline rate as that of other traders?

Use Monte Carlo simulation (with at least 100000 simulations) to calculate a p-value under this null hypothesis. Include the following items in your write-up:

- the null hypothesis that you are testing;
- the test statistic you used to measure evidence against the null hypothesis;
- a plot of the probability distribution of the test statistic, assuming that the null hypothesis is true;

- the p-value itself;
- and a one-sentence conclusion about the extent to which you think the null hypothesis looks plausible in light of the data. This one is open to interpretation! Make sure to defend your conclusion.

Problem 4: milk demand, revisited

Return to the `milk.csv` data set that we analyzed in class. Recall that these data are based on a stated preference study on consumers' price sensitivity for milk; there are two variables of interest here:

- **price**, representing the price of milk
- **sales**, representing the number of participants willing to purchase milk at that price.

Your task is to use bootstrapping to quantify your uncertainty regarding the **price elasticity of demand** for milk based on this data. For this problem you should turn in a single figure showing the bootstrap sampling distribution for the elasticity parameter. This figure should have an informative caption in which you explain what is shown in the figure and also quote a 95% bootstrap confidence interval for the elasticity, rounded to two decimal places. Use at least 10,000 bootstrap samples to produce your figure and confidence interval.

Problem 5: standard-error calculations

Part A. Suppose that $X_1, \dots, X_N \sim \text{Bernoulli}(p)$ and that $Y_1, \dots, Y_M \sim \text{Bernoulli}(q)$ (all independent). We will consider $\hat{p} = \bar{X}_N$ and $\hat{q} = \bar{Y}_M$ as estimators of p and q , respectively.

- Show that $E(\hat{p} - \hat{q}) = p - q$, the true difference in success probabilities.
- Use what you know of probability to compute the standard error of \hat{p} , i.e. the standard deviation of the sampling distribution of \hat{p} .
- Compute the standard error of $\hat{\Delta} = \hat{p} - \hat{q}$ as an estimator of the *true* difference $\Delta = p - q$.

Part B. Suppose we have data on some numerical attribute from two groups: X_1, \dots, X_N from group 1, and Y_1, \dots, Y_M from group 2. (Notice the unequal sample sizes N and M .) Suppose that:

- $E(X_i) = \mu_X$ and $\text{var}(X_i) = \sigma_X^2$, both unknown
- $E(Y_i) = \mu_Y$ and $\text{var}(Y_i) = \sigma_Y^2$, again both unknown

Suppose we're interested in the population-level difference in means between the two groups: $\Delta = \mu_X - \mu_Y$. We use the difference in sample means to estimate this quantity:

$$\hat{\Delta} = \bar{X}_N - \bar{Y}_M$$

Use your knowledge of probability theory to calculate the expected value and standard error of $\hat{\Delta}$. These expressions should involve some combination of the true unknown population parameters and the sample sizes.