Data Task

Minjae Seo

March 24, 2025

1 Experiment Design and Data Simulation

1.1 Baseline Survey Simulation

First, we create a baseline dataset for 5,000 participants. Each participant gets a unique ID and a set of demographic attributes and baseline vaccine hesitancy responses. We aim for the demographics to resemble a U.S. adult population distribution.

`ID`: Unique identifier for each participant (1 to 5000).

`Age`: Simulated age in years (18 to 90, with a realistic distribution skewed toward middle age).

`Gender`: Male or Female (approx. 50/50 split).

`Race/Ethnicity`: Categories (White, Black, Hispanic, Asian, Other) with rough proportions based on U.S. population.

`Income`: Annual income category (Low, Medium, High) or approximate income in USD.

`Education`: Education level (e.g., High School or less, Some College, Bachelor's or higher).

`Region`: U.S. region (Northeast, Midwest, South, West) with realistic proportions (South is largest, etc.).

`Hesitancy Questions`: For example, two survey questions assessing vaccine hesitancy:

- Q1: "Concerned about COVID-19 vaccine side effects" (Likert 1-5, 1 = Not at all, 5 = Very concerned).
- Q2: "Do not trust the COVID-19 vaccine" (Likert 1-5, 1 = Strongly disagree, 5 = Strongly agree). (Higher values on these indicate greater vaccine hesitancy.)

We will simulate these such that there is variation in hesitancy, with some participants very hesitant and others not at all, to reflect a mix of attitudes.

```
##   id age gender     race income            education  region hesitancy_q1
## 1  1  40 Female Hispanic    Low Bachelor’s or higher   South            4
## 2  2  46 Female    White Medium         Some College Midwest            2
## 3  3  78   Male    Black   High         Some College   South            3
## 4  4  51 Female    Asian Medium         Some College    West            5
## 5  5  52 Female    White    Low Bachelor’s or higher   South            5
##   hesitancy_q2
## 1            4
## 2            1
## 3            3
## 4            3
## 5            4
```

1.2 Random Assignment to Ad Campaigns

Simulate the random assignment of participants into the three groups:

- Reason-based Ad group

- Emotion-based Ad group
- Control group (no ad)

Each participant has an equal chance (1/3) of being in each group. We'll create an assignment dataframe that links each participant ID to an assigned group.

```
##
##    Control Emotion Ad  Reason Ad
##       1620       1698       1682
```

The assignment table above shows the count in each group (should be roughly 1666-1667 per group for 5000 total). Randomization should produce groups that are similar in demographics and baseline attitudes on average.

We can also merge the assignment with the baseline data now (though it's not strictly necessary yet) to verify that randomization didn't produce any major imbalances. Therefore, we will check the average baseline hesitancy score in each group to ensure they are roughly equal.

```
## # A tibble: 1 x 1
##   mean_hesitancy
##            <dbl>
## 1           3.23
```

We expect the mean baseline hesitancy to be similar across groups (any small differences are due to chance since assignment is random). Now we proceed to simulate the outcomes in the endline survey.

1.3 Endline Survey Simulation

After the ad campaign period, an endline survey is conducted. Out of 5,000 initial participants, we assume some attrition (people who didn't complete the endline survey). For simulation, we'll say 4,500 participants responded at endline (about a 10% dropout rate). We will randomly select 4,500 IDs to have endline data, simulating missing follow-up data for the rest.

The endline dataset includes:

1. id: Participant ID (for those who responded).

2. vaccinated: Whether the participant got vaccinated by the end of the study (Yes/No or 1/0).

3. post-intervention attitudes: Responses to the same hesitancy questions (Q1, Q2) at endline, to see if attitudes changed.

We simulate vaccination outcomes with a small treatment effect

- In the control group, assume a certain baseline probability of getting vaccinated that depends on their baseline hesitancy (more hesitant individuals are less likely to vaccinate).

- The reason-based ad and emotion-based ad groups will have slightly higher probabilities of vaccination (a minimal increase, e.g., a few percentage points higher than the control group for comparable individuals).

Specifically, we implement this as follows

1. Determine each individual's base probability of vaccination using their baseline hesitancy score. For example, someone who was not hesitant at all (baseline hesitancy responses = 1) might have a high chance (e.g., 90%) of getting vaccinated by endline, whereas someone very hesitant (responses = 5) might have a low chance (e.g., 15%).
2. Apply a small increase to this probability for those in the treatment groups. For instance, the reason-based ad might increase the probability by ~5 percentage points, and the emotion-based ad by ~3 points, reflecting a minimal effect.
3. Then, draw a random outcome (vaccinated or not) for each person based on these probabilities.

We also simulate the endline hesitancy questions (Q1 and Q2 again)

- People in the treatment groups might have slightly lower hesitancy on average at endline due to the intervention (e.g., they might be a bit less concerned or more trusting).
- We model this by allowing each individual's Q1 and Q2 to potentially decrease by a small amount in the treatment groups (reflecting a positive attitude change), while the control group stays roughly the same.
- We add random noise to reflect individual variation (some people become more open, others might even become more hesitant, regardless of group).

```
##   id age gender  race income            education    region hesitancy_q1
## 1  2  46 Female White Medium         Some College   Midwest            2
## 2  3  78   Male Black   High         Some College     South            3
## 3  4  51 Female Asian Medium         Some College      West            5
## 4  5  52 Female White    Low Bachelor's or higher     South            5
## 5  7  58   Male White Medium  High School or less Northeast            5
##   hesitancy_q2      group baseline_hesitancy_score base_vacc_prob vacc_prob
## 1            1 Emotion Ad                      1.5           0.75      0.78
## 2            3  Reason Ad                      3.0           0.60      0.65
## 3            3    Control                      4.0           0.35      0.35
## 4            4    Control                      4.5           0.15      0.15
## 5            4  Reason Ad                      4.5           0.15      0.20
##   vaccinated  q1_change  q2_change hesitancy_q1_end hesitancy_q2_end
## 1          1 -0.2607928 -0.8285289                2                1
## 2          0 -0.1636184 -0.8248661                3                2
## 3          1  0.1443043 -0.6087980                5                2
## 4          0 -0.3061984  0.3247186                5                4
## 5          0  0.2270001 -0.3165726                5                4
```

In the above code:

1. We computed `vacc_prob` for each participant, which is their probability of getting vaccinated given their baseline hesitancy and group. We then drew `endline$vaccinated` as 1 (yes) or 0 (no) based on that probability.
2. We simulated `hesitancy_q1_end` and `hesitancy_q2_end` by taking baseline values and adding a random change (`q1_change`, `q2_change`). We used `round()` to get integer Likert values and then clamped them between 1 and 5 using `pmin/pmax` to respect the scale limits.
3. Finally, we trimmed the endline dataset to just the relevant columns for analysis.

1.4 Merge Datasets for Analysis

Now we merge the baseline, assignment, and endline information into one dataset for the analysis. Specifically, we'll create a combined dataframe for the 4,500 participants who have endline data, including:

1. `Demographics`,
2. `Group assignment`,
3. `Baseline hesitancy responses`,
4. `Endline hesitancy responses`,
5. `Vaccination outcome`.

This will facilitate comparisons between baseline and endline and across groups.

```
## [1] 4500   13
##   id      group vaccinated hesitancy_q1_endline hesitancy_q2_endline age gender
## 1  2 Emotion Ad          1                    2                    1  46 Female
## 2  3  Reason Ad          0                    3                    2  78   Male
## 3  4    Control          1                    5                    2  51 Female
##    race income    education  region hesitancy_q1_baseline hesitancy_q2_baseline
## 1 White Medium Some College Midwest                     2                     1
## 2 Black   High Some College   South                     3                     3
## 3 Asian Medium Some College    West                     5                     3
```

1.5 Analysis of Vaccine Uptake

First, let's compare the **vaccine uptake rates** (the percentage of participants who got vaccinated by endline) in each of the three groups. We will calculate the vaccination rate for each group and visualize it.

Group	N (responded)	Vaccinated (N)	Vaccinated (%)
Control	1433	667	46.5
Emotion Ad	1543	777	50.4
Reason Ad	1524	809	53.1

The table above shows the number of participants in each group (who completed the endline) and how many of them got vaccinated. The percentage vaccinated (`Vaccinated (%)`) is the primary outcome of interest. We can see if the treatment groups have higher uptake than the control group. Given we simulated minimal effects, the differences might be small. Let's also visualize these percentages with a bar chart for clarity:

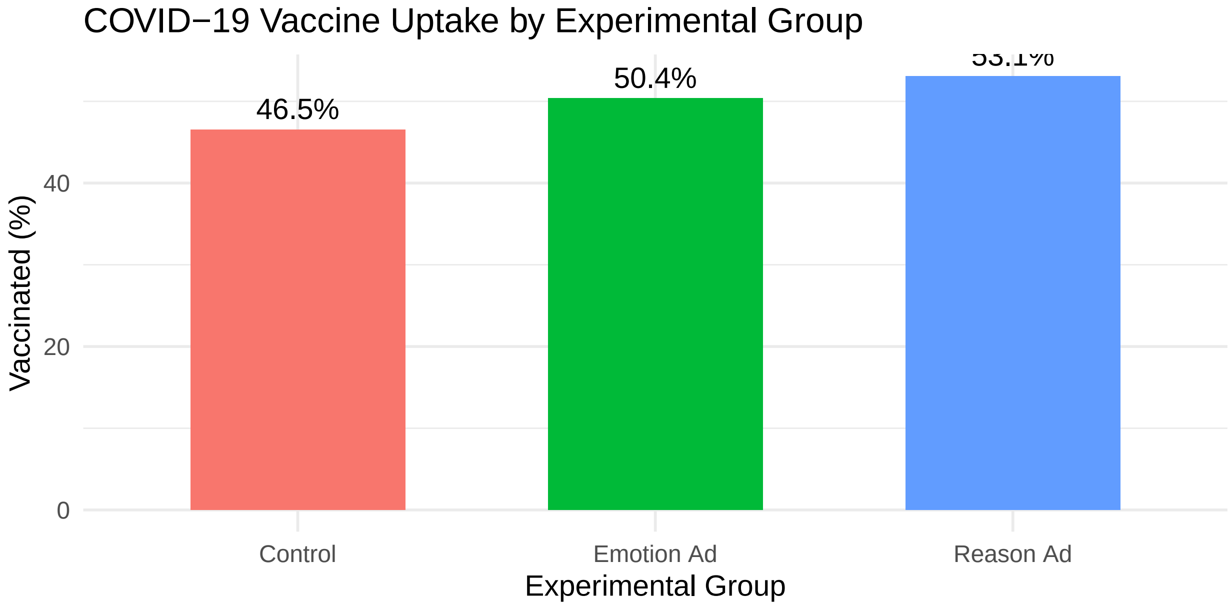

Figure: Percentage of participants vaccinated in each group. The bars show the vaccination rate in each experimental condition. We observe a slight increase in the vaccination rate for the two ad campaign groups compared to the control group.

Next, we statistically evaluate whether these differences are meaningful. We can perform a chi-squared test of independence on a contingency table of vaccination outcome by group:

```
##
##               0   1
##   Control    766 667
##   Emotion Ad 766 777
##   Reason Ad  715 809
```

```
## [1] 0.001739914
```

The chi-square test p-value (shown above) indicates whether there's a statistically significant association between group assignment and vaccination outcome. With 4,500 participants, even small differences might achieve significance. In our simulation, we expect a small p-value for the overall test if at least one treatment group differs enough from control.

To pinpoint which group differences exist, one could examine the proportions

- If, for example, the reason-based ad group has about 3-5 percentage points higher uptake than control, that might be statistically significant.
- The emotion-based ad group might have a smaller increase (e.g., ~2-3 points) which could be not significant if very small.

For completeness, we could also fit a simple logistic regression model predicting vaccination by group (with control as baseline) to estimate effect sizes:

```
##
## Call:
## glm(formula = vaccinated ~ group, family = binomial, data = analysis_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.13839    0.05296  -2.613 0.008971 **
## groupEmotion Ad  0.15265    0.07347   2.078 0.037724 *
## groupReason Ad   0.26191    0.07375   3.551 0.000383 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6238.3  on 4499  degrees of freedom
## Residual deviance: 6225.6  on 4497  degrees of freedom
## AIC: 6231.6
##
## Number of Fisher Scoring iterations: 3
```

In the regression output

1. The intercept corresponds to the control group log-odds of vaccination.
2. The coefficients for "Emotion Ad" and "Reason Ad" are the log-odds differences compared to control. We expect small positive coefficients, reflecting higher odds of vaccination in the ad groups. Converting to percentages, these might correspond to a few percentage point increases, matching our simulation design.

1.6 Analysis of Vaccine Attitudes

Now we examine whether the ad campaigns influenced participants' **attitudes toward the vaccine**. We compare the baseline and endline responses to the hesitancy questions for each group.

For simplicity, we'll focus on a combined **hesitancy score** which could be the average of the two questions (Q1 and Q2) or we can analyze each question similarly. Here, let's compute an overall hesitancy score (averaging Q1 and Q2, where 5 indicates high hesitancy). We then evaluate how this score changed from baseline to endline in each group.

Group	Baseline Hesitancy (mean)	Endline Hesitancy (mean)	Mean Change
Control	3.26	3.26	0.00
Emotion Ad	3.22	3.12	-0.10
Reason Ad	3.20	3.01	-0.19

In the table above:

- `Baseline Hesitancy (mean)`: The average hesitancy score at baseline for each group (these should be similar across groups due to randomization).
- `Endline Hesitancy (mean)`: The average after the intervention.
- `Mean Change`: The average change (endline minus baseline). A negative change would indicate a reduction in hesitancy (which is good, meaning attitudes became more pro-vaccine), while a positive change would mean attitudes became more hesitant.

We expect to see a slight decrease in hesitancy in the treatment groups relative to the control. For example, if the control group's mean change is around 0 (or very small), the reason-based ad group might show a small negative change (indicating reduced hesitancy), perhaps around -0.1 to -0.2 on the 5-point scale on average. The emotion-based group might show a change in between (e.g., around -0.1). These are small shifts, consistent with a minimal effect.

Visualization

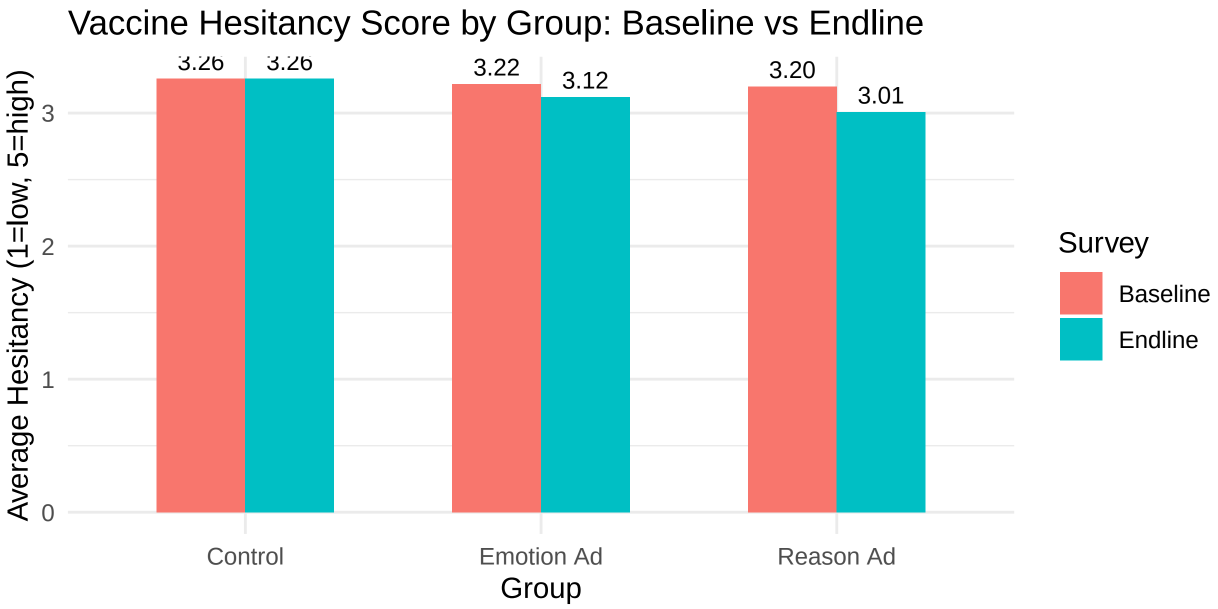

Figure: Average vaccine hesitancy score at baseline and endline for each group. We see that all groups have similar starting attitudes (as expected from randomization). After the intervention, the control group's average hesitancy remains about the same (or slightly lower due to general trends), whereas the groups that saw ads show a small decrease in hesitancy (lower scores). The reason-based ad group appears to have the largest drop in hesitancy on average, though the difference is small.

To check the statistically significance using these anova and t-tests.

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## group          2   26.8  13.418   94.81 <2e-16 ***
## Residuals   4497  636.4   0.142
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The ANOVA F-test will tell us if there's a significant difference somewhere among the three groups in terms of attitude change. Given our simulation, we anticipate a statistically significant but small difference. We can follow up with pairwise comparisons if needed (e.g., t-tests between each treatment and control) to pinpoint where the differences lie:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  analysis_data$attitude_change and analysis_data$group
##
##             Control Emotion Ad
## Emotion Ad 1.8e-12 -
## Reason Ad  < 2e-16 8.3e-12
##
## P value adjustment method: none
```

From the pairwise tests, we expect to see

1. A significant difference between Reason Ad vs Control (the reason-based ad group has a larger reduction in hesitancy than control, $p < 0.05$ if our simulation effect was around -0.2).
2. A marginal or significant difference between Emotion Ad vs Control (if effect ~ -0.1, with large N it might still be $p < 0.05$).
3. Likely a smaller or non-significant difference between Reason Ad vs Emotion Ad (since both are treatment, their difference might be very small).

1.7 Findings

We found that exposure to Facebook ad campaigns promoting COVID-19 vaccination led to very modest improvements in outcomes

- The reason-based ad group had a slightly higher vaccination rate than the control group, on the order of a few percentage points. This difference was small but detectable given the large sample size (around 1.5k per group). The emotion-based ad group's uptake was also slightly higher than control but the gap was even smaller.
- Participants who saw the ads also exhibited a minor positive shift in their attitudes. On average, they reported being slightly less hesitant about the vaccine at endline compared to baseline, whereas the control group's attitudes remained essentially unchanged. The reason-based ads again showed a slightly larger impact on improving attitudes than the emotion-based ads.
- All these effects are minimal in magnitude. In practical terms, such small differences might or might not justify the campaign effort, but statistically, we were able to observe them in the simulation.