W241 | Summer 2018 | Section 1

Progress Report

Jack Workman and Yulia Zamriy

Main research question

To what degree, if any, do reviews influence an individual's perception of a short story?

Motivation

Online reviews have become a strong force in determining business success. If you want to pick a restaurant for dinner, it is quite likely that you will check a few on Google Maps or Yelp to get other people's opinions. If you need to buy something on Amazon and there are multiple options, average score will probably be one of the main factors in your decision. What new movie should you watch tonight? It depends on Rotten Tomatoes score or New York Times review.

However, the power of online reviews does not stop at the decision point. While eating a meal in a 5-star restaurant, will we persuade ourselves that it is worth all five stars even if it is not? After buying a standard plastic storage bin on Amazon with a review score of 3.5 (because somehow perfectly-reviewed bins do not exist), are we going to consider it good-enough but not great despite it being completely functional? If our friends make us watch a poorly reviewed movie, will we find reasons to justify the low score and ignore the good parts of the film?

The hypothesis we would like to test is whether we are so reliant on strangers' opinions that it is hard to form our own independent views about consumed products. If we find enough evidence to support this hypothesis, the implications for businesses are considerable. Boosting product reviews would not only drive short-term product sales, but also might help with repeat purchases from "satisfied" consumers.

However, if we find no evidence of consumer compliance to public opinion, it might give us a glimpse of hope that we still can be independent thinkers.

Experiment design

Null hypothesis: average review score of the short story is not relevant to the respondent's review.

Alternative hypothesis: average review score of the short story is a significant factor of determining respondent's review (the higher the average review the higher the respondent's score and vice versa).

The experiment will be conducted via an online survey that will consist of the following:

- Brief explanation of what to expect
- Short story (about 5-minute read)
- Page to rate the story
- Question to determine if the responder is familiar with the story
- 3-4 questions to collect some demographic data (gender, age, education level etc.)

The treatment will consist of prominently displayed average rating of the story. There are four potential treatment groups:

1. Control: no average review provided. The users are supposed to rate the story without any external suggestions
2. Treatment #1: display mediocre rating (for example, 2.5 on a 5-point scale)
3. Treatment #2: display high rating (for example, 4.5 on a 5-point scale)
4. Treatment #3: display low rating (for example, 1 on a 5-point scale)

The story will be the same across all groups. It is relatively obscure to avoid bias and limit the chances of experiment participants being familiar with the short story. Moreover, this story does not contain any references to current affairs and received a 4.9/7 rating on a hosting website, both of which should prevent study participants from forming strong opinions about it.

Subject recruitment will be conducted via two sources:

1. Personal connections (Facebook, email, coworkers etc.)
2. Amazon Mechanical Turk

Random assignment of recruited subjects will be performed using Qualtrics Randomizer capabilities.

Results Analysis

Metrics of interest:

- Primary: average rating
- Secondary: is there correlation between user rating and time spent on the survey?

Most likely we'll need to perform blocking by the source of the participants (personal connections vs. Mechanical Turk). There are no obvious reasons why the results would be different across the two sources. However, it is possible that:

- Subjects from Mechanical Turk are more familiar with this type of study and, hence, might be considered "primed"
- Personal connections might skew towards specific personality types

Tools

We are planning to use the following tools :

- Qualtrics to build and serve the survey (as well as participant assignment to treatment groups)
- Mechanical Turk for participant recruitment
- Google Sheets for response tracking and timeline management

Labor division

As we are still early in the planning stage, it is hard to assign specific roles to project members. However, here is a preliminary list of tasks that we will need to distribute:

- Designing survey
- Writing a standard message for volunteer recruitment
- Building an email list
- Managing Mechanical Turk
- Tracking responses
- Processing data

- Analyzing data
- Building presentation
- Writing the report

Outstanding questions for the project members to research

1. Should we create two versions of each survey to track Mechanical Turk vs. personal contacts responses? Or is there a way to track referring source within chosen survey platform?
2. Should we include a few mock text reviews to support scale rating? If so, how many?
3. What is the expected budget for Mechanical Turk?
4. How should we compose volunteer recruitment message? How do we motivate people to participate without biasing them? Should the message be personalized or generic?
5. Should we ask respondents to write a brief review?

Questions for the instructor

1. How many treatment groups can we allow? How many participants should we aim for in each group? Potentially, we could estimate number of responders based on the compiled list of personal connections and Mechanical Turk background information.
2. How wide should the rating scale be? Is 5-point scale too narrow? Is 10-point scale too wide?
3. What kind of demographic questions to include? Should we included these questions at the end or at the beginning of the survey?