PHP2550: Practical Data Analysis

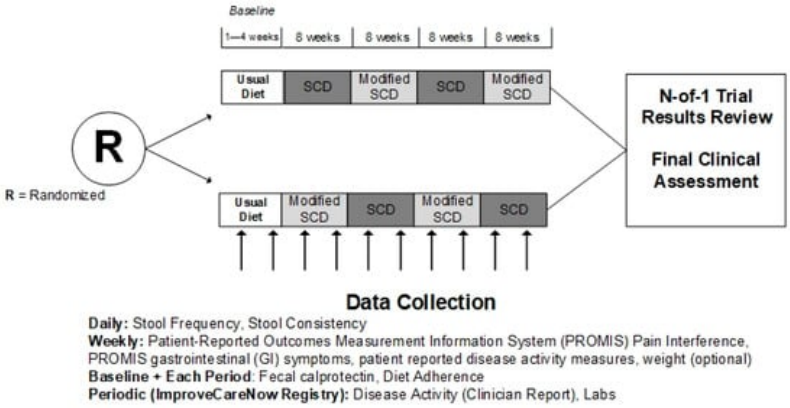
Assignment 1: Exploratory Data Analysis

Antonella Basso

September 30, 2022

1. Data Visualization and Reports

Read the paper *Personalized Research on Diet in Ulcerative Colitis and Crohn's Disease: A Series of N-of-1 Diet Trials* by Kaplan et al (available on Canvas as "diet_trials.pdf"). The image below gives a summary of how data was collected for each individual. We will focus on how the data and results are presented so don't get hung up on understanding all pieces of the statistical analysis. Respond to the following prompts about the paper.

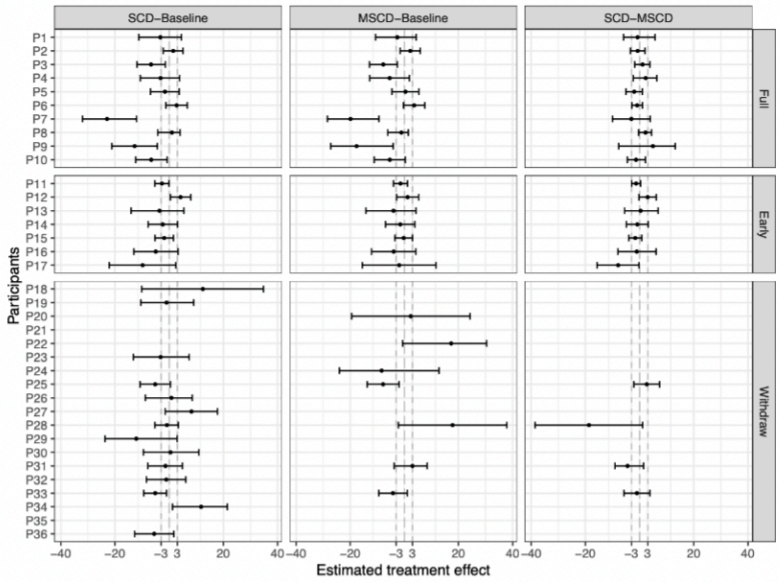


a. Summarize the paper in 2-3 paragraphs using your own words (a good way to help this is to not look at the paper when writing your response). What interested you about the paper? Were there any unfamiliar terms or methodology?

b. Reflecting on the replication crisis and documentation discussion from class, how easy would it be to try to replicate this study? Explain your response.

c. Think about the audience of this paper and look at how the data collection and data were presented. Compare this to the paper on sweetened beverages.

d. Take a look at the visual below that did not make into the paper and was replaced with Figure 1 in the paper. How is the data presented in each visual? Why do you think the authors went with Figure 1?



Solution

a. The Personalized Research on Diet in Ulcerative Colitis and Crohn's Disease (PRODUCE) study, conducted a series of double-crossover N-of-1 trials (54 single-subject) to test the comparative effectiveness of two grain-free, low-sugar, nutritionally complete diets—the specific carbohydrate diet (SCD) and the more liberal, modified version, MSCD—in children and adolescents (aged 7-18 years) with either ulcerative/indeterminate colitis (UC/IC) or Crohn's disease (CD) that suffer from inflammatory bowel disease (IBD) symptoms and active inflammation. Specifically, patients were recruited from 19 sites within the learning health network ImproveCareNow (ICN) and randomized to "1 of 2 sequences of 4 alternating 8-week SCD and MSCD periods", after a 2-week period under their usual diets (UD) for baseline comparison. Although no single outcome of interest was specified to give providers the freedom of deciding which measures are most relevant, the primary study outcomes included: fecal calprotectin obtained from stool samples (at baseline and once per diet period); daily stool consistency and frequency; and weekly IBD symptoms and pain interference—all of which were in the form of raw patient-reported outcomes (PROs). Results from the study showed that there was no clinically meaningful difference between the diet therapies for most individuals. They do suggest however, that SCD and MSCD might improve IBD symptoms and inflammation in some children and adolescents when compared to a usual diet. Moreover, despite having demonstrated high attrition and difficulty with participant retention (only 39% of all participants fully completed the trial), not only did the majority of responders remain on one of the alternative diets after trial completion, but many families and providers felt that they gained adequate information from the N-of-1 trial to guide shared decision-making. Notably, the study was conducted in efforts to expand research around the role of dietary therapy in pediatric IBD, inform individuals about the potential benefits of these interventions, and allow for personalized treatment decisions for those suffering from IBD.

   One notable approach taken by researchers in this study, despite my being unfamiliar with the specific (combined) methodology, was the use of Bayesian generalized linear mixed models with noninformative priors to aggregate individual results and obtain average treatment effects, using noninformative prior distributions. Presumably, the choices for noninformative prior distributions were made on account of both the limited evidence of specific carbohydrate diets for IBD, and the expected (and confirmed) heterogeneity of treatment effects across individuals. According to researchers, such models "generated efficacy data and individualized probabilistic estimates of effectiveness for each diet that enabled

participants to make personalized decisions about using diet to manage IBD". Moreover, the addressed limitations regarding the study's (lack of) generalizability were particularly compelling in suggesting possible directions for future research. For example, having primarily enrolled patients with CD, it may be the case that pooled estimates from aggregate analyses (which include all patients regardless of disease type) are not only biased with respect to the majority group, but "may not be as generalizable to patients with UC/IC" in being more representative of patients with CD. Additionally, the fact that "the study population was primarily white and insured" also drastically limits the generalizability of these findings in that it narrows down the target population to a much smaller socioeconomic group with greater access to (personalized) treatments. Considering these limitations in future studies may yield more robust effect estimates and broaden the study's relevance in benefiting a broader population that demands representation in clinical research.

b. Excluding the disclaimer which clearly states that "unauthorized reproduction of this article is prohibited", replicating the study itself may be possible given that methodology is specifically outlined, and patient characteristics as well as trial results and corresponding summary statistics are given by the tables in the article or by the linked "supplementary digital content". However, the fact that no code is available and that some relevant of the patient-specific data is summarized by figures or statistics would make it almost impossible to conduct similar analyses. Indeed some aspects of the study could be reproduced, but others would require additional information to attempt even roughly. In either case, replicating any methods proposed in the study would be both tasking and time-consuming. This inability to follow the study's analyses and procedures efficiently (if even effectively) leads us to believe it would be very difficult to replicate.

c. The methodological rigor and terminology presented in this study suggests that its most likely target audience consists of medical experts, clinicians, and statistical or public health researchers; particularly those interested in dietary therapies and/or pediatric IBD. Although the analyses conducted here may be of particular interest to such academic experts, the study's findings themselves, which are interesting, informative, and easily interpretable, leads us to believe that it may also be appealing to family or patients suffering from CD or UC/IC. While the paper on sweetened beverages may similarly appeal to a greater (not-necessarily-academic) population, it differs in being centered around public health policy, and hence, targeting a different class of researchers and (bio-)statisticians nonetheless. The ways in which both studies were presented graphically, may be positive for those who may be more familiar with statistical tables and graphs or those who are able to gain more intuition visually. However, the fact that this article included more graphs that were more detailed and complex makes the prospect of non-academics gaining sufficient information from them very unlikely.

d. Both visuals are intended to give a general overview of the relative effects of each diet treatment for each individual N-of-1 trial—segmented by participant status (i.e., full completers, early completers, and withdrawals) and particular diet comparison (i.e., (a) SCD vs. UD, (b) MSCD vs. UD, and (c) SCD vs. MSCD). Specifically, each graph was divided into three sub-sections in each direction (vertical and horizontal) to give nine "smaller graphs" of estimated treatment effect for each combination of patient status and diet comparison. Notably, the visual that was included in the article is far more intricate that the one above in several respects. First, in that Figure 1 in the article represents "estimated treatment effects" as probabilities (0-1) of symptomatic improvement in the *Pediatric IBD Symptom Scale* (0-4)—a scale that has not been verified—making the interpratation of values far more complex and counterintuitive. Second, in that the graph chosen for the article splits each of the nine sub-sections into two smaller groups to further break up individuals by a disease-type factor (CD vs. UC/IC), resulting in a total of 18 "sub-graphs" as opposed to 9. Third, in that the manner the data are displayed and represented in the article's graph is far less straight-forward that in the more simplistic boxplot-like visuals in the graph above. Particularly, while the above plot uses what we presume to be ranges with a marker for central tendency (either mean or median), the other uses a color scheme within bar graphs, which requires additional effort to decipher even with a detailed description. While the graph not included in the article is superior in being simplistic and easily interpretable, it lacks in providing both less (or more generalized) information (regarding changes in IBD symptoms on a relative basis), and no descriptions or way of confirming what the graph actually aims to convey.

2. Exploratory Analysis

The file `ssm_survey.csv` contains survey data from two studies. Participants in each study were Southern California residents who registered to vote in precincts that supported a ballot measure banning same-sex marriage in 2008. These voters were then recruited to participate in an online survey panel about politics and randomly assigned to a treatment. The first group was assigned to receive the same-sex marriage script from a gay canvasser. The second group was assigned to receive the same-sex marriage script from a straight canvasser. Groups three and four were encouraged to recycle household waste by gay or straight canvassers, respectively; however, canvassers did not reveal their sexual orientation when delivering the recycling script. The fifth group was a control group to which no canvassers were assigned. In the second study, participants were only selected to receive the same-sex marriage script from a gay canvasser or no contact. The first study started with the initial survey at the end of May 2013 and the second study in August 2013. After canvassing, additional responses to this survey were collected at 3, 12, 23, 27, 45, and/or 280 days after contact. These are referred to as the waves of the survey. The variables in the data set are listed below and are described using the language used in the study.

- `panelist_id`: unique id for each recruited participant
- `study`: study id (Study 1 or Study 2)
- `treatment_assignment`: {C, R by G, R by S, G by G, G by S} either C (no contact) or representing which script (R = recycling, G = same-sex marriage) and sexual orientation of the canvasser (S = straight, G = gay)
- `wave`: wave of the online survey (1 to 7)
- `therm_level`: response to a feeling thermometer question about gays and lesbians with 0 = unfavorable and 100 = favorable
- `therm_change`: change in therm level from last measurement

Conduct an initial exploratory analysis of this data. As part of your analysis, you should consider the following questions: what differences are there between the two studies initial conditions and results? how did the participants' therm level change over time within each treatment group?

Solution

Considering the large quantity of missing values in the data, noticeable at first glance, we first attempt to quantify this missingness as well as identify possible reasons behind it. Upon closer inspection, we notice that there are exactly 34,675 values missing from each of the `therm_level` and `therm_change` columns. Having found that these correspond to the same rows is suggestive of both a dependency between the two variables and that missingness likely results from participants not having provided responses for all 7 waves. In fact, we find that all participants had at most 5 observations—Table 1 below provides the number of participants (out of a total 11,948 in both studies) for whom 1, 2, 3, 4, and 5 responses were observed.

Table 1: Number of Observed Responses for Participants

Observations	Participants (n)
1	343
2	3388
3	6225
4	837
5	1155

As is also reflected in Figure 4, not only do we see no responses corresponding to waves 5 and 6, but Study 2 is the only study to have observed more than 3 responses for each participant. That is, while Study 1 only observes panelist responses for the first, second, and seventh waves, Study 2 obtains responses for the first four and last (seventh) waves. Restricting our attention to only the observed cases for this exploratory data analysis (EDA), we proceed on the subset of the data with no missing values.

Table 2 below provides summary statistics (count, mean, and standard deviation) for each study and treatment assignment combination in terms of thermometer level (TL) and thermometer change (TC). Moreover, Figure 1 below provides an overview of the observed TL densities in each of the studies. Noticeably, both are nearly identical. This is surprising given that Study 1 includes the effects of 3 additional "treatments". However, this could be, in part, due to greater influence on the first study's data by its majority treatment groups ("C" and "G by G") which are the same as those of the second study's, and/or the notion that panelists' opinions are likely to remain pretty consistent across all views of the opposing political party.

Table 2: Descriptive Statistics of Thermometer Level (TL) and Thermometer Change (TC)

			TL		TC	
Study	Treatment	Participants (n)	Mean (μ)	SD (σ)	Mean (μ)	SD (σ)
Study 1	C	13516	59.8466	27.8777	1.2473	9.1786
Study 1	G by G	2965	60.0169	28.1149	2.2597	9.5358
Study 1	G by S	2654	60.3282	26.8368	1.4687	9.4767
Study 1	R by G	2706	59.4409	28.5885	1.0281	9.3018
Study 1	R by S	2691	59.2490	28.1166	1.3382	10.0580
Study 2	C	5098	58.0530	28.2027	0.2823	8.4138
Study 2	G by G	5287	61.5792	27.8765	2.3291	9.8969

Fig.1: Thermometer Level Densities by Study

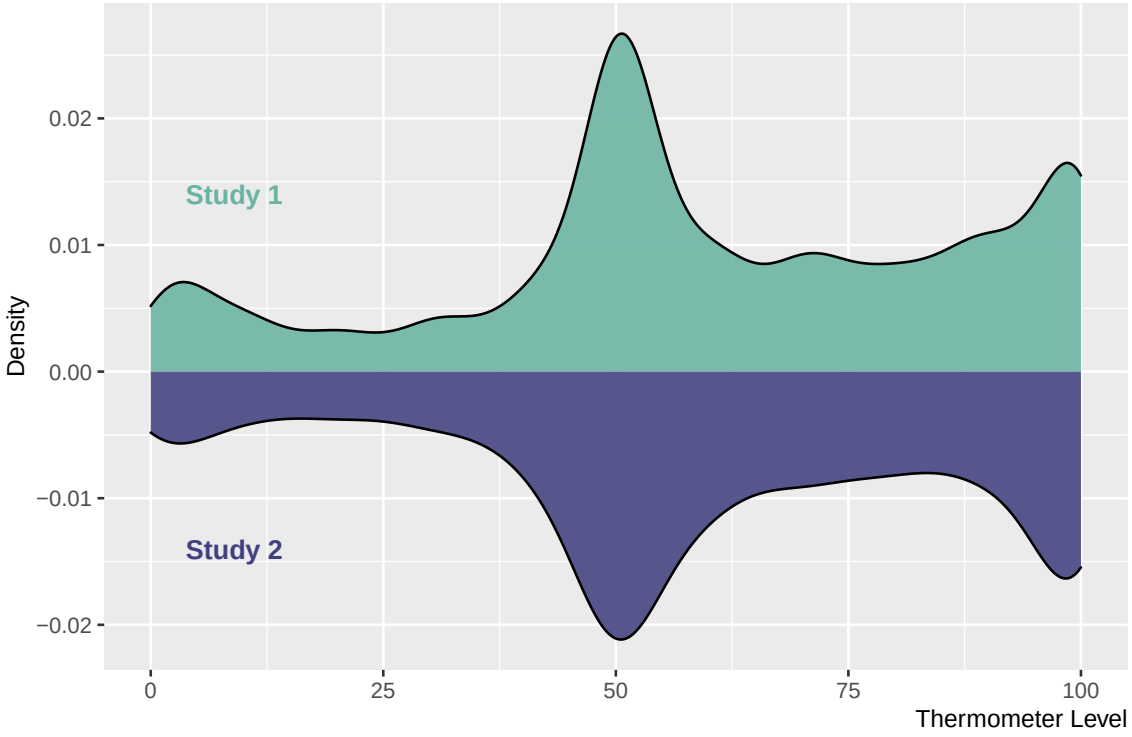
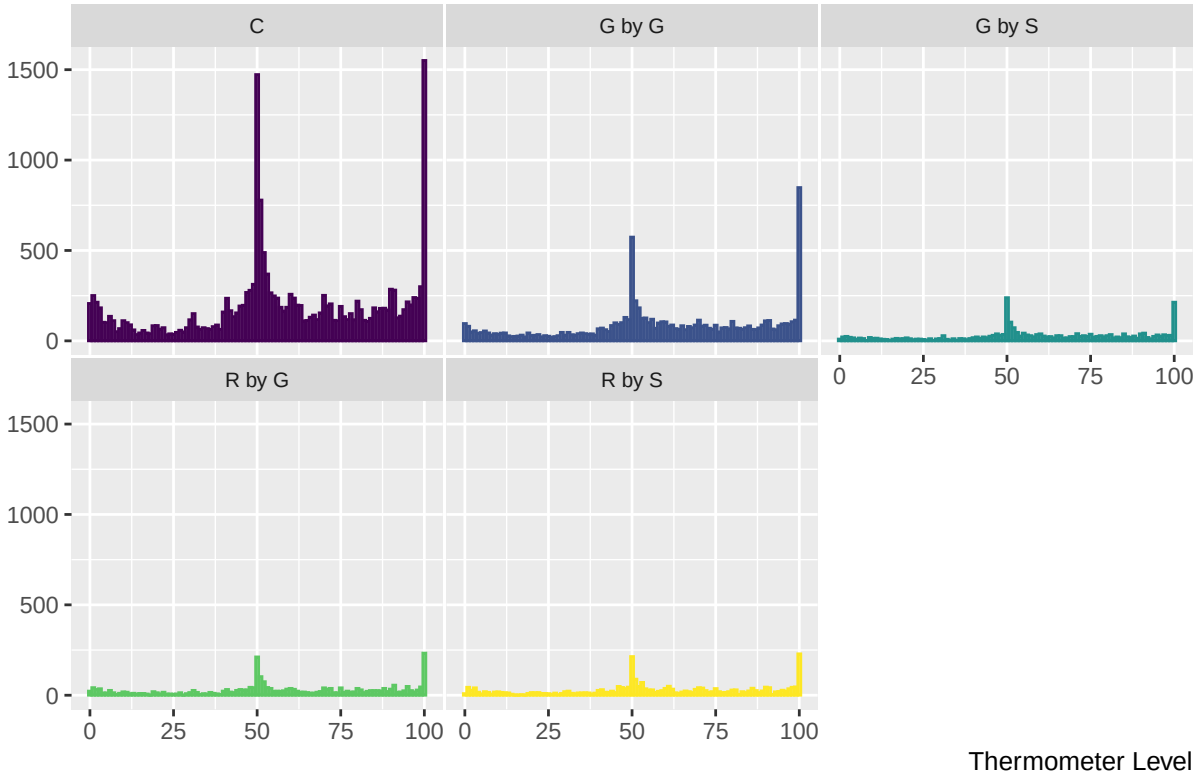
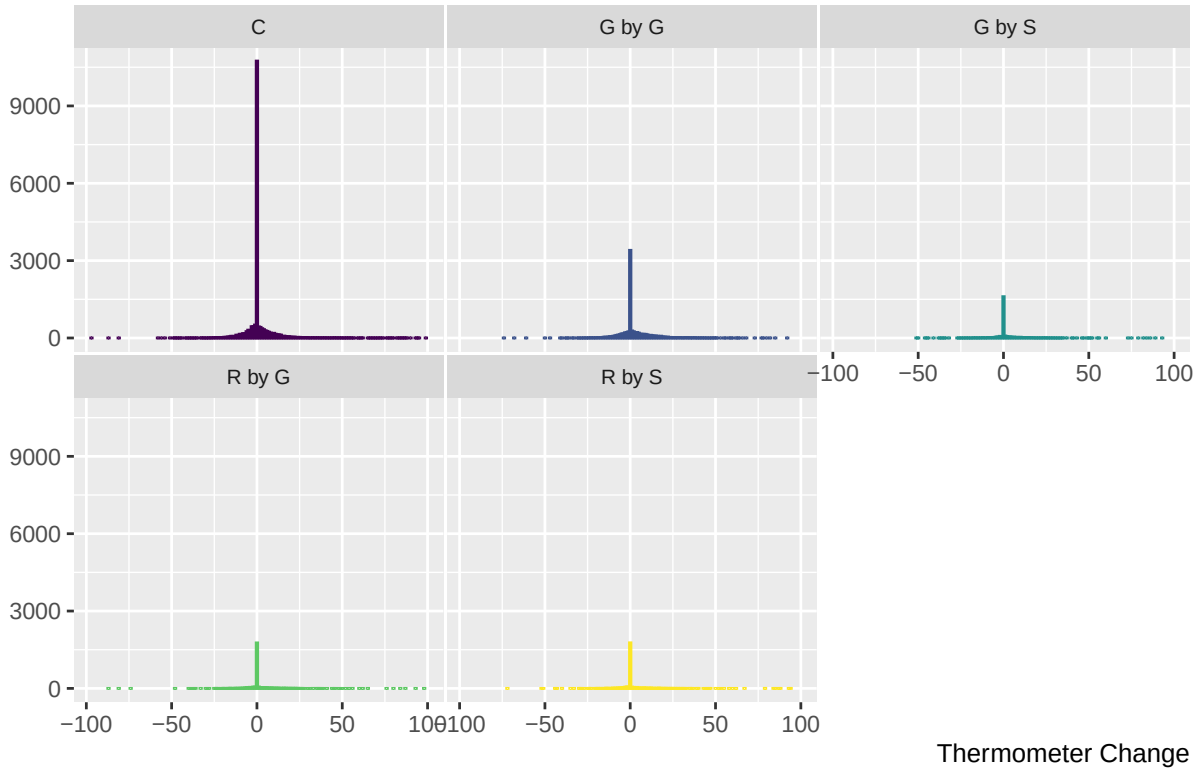


Fig.2: Frequency Distributions of TL and TC by Treatment Assignment



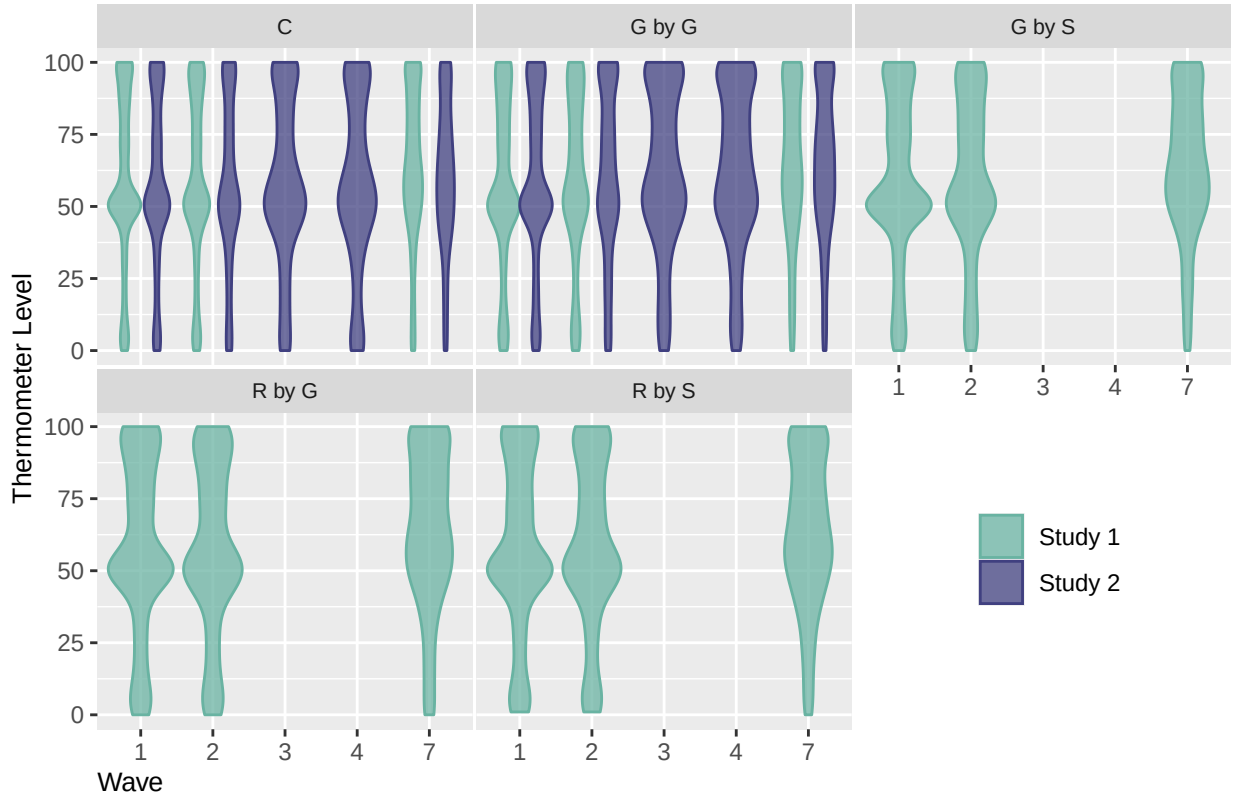
Thermometer Level



Thermometer Change

We notice similar patterns in Figures 2 (above) and 3 (below), which provide less generalized overviews of the observed data than Figure 1; specifically, in focusing on how the data is distributed in each treatment group. Figure 3 for example, shows us that thermometer levels are largely concentrated around mean values of ≈ 50. It should be noted however, that this more narrow representation of the data shows us that the positive tails of treatment-specific distributions become progressively wider with increasing time (that is, later "waves").

Fig.3: Spread of Thermometer Level by Study, Treatment Assignment, and Wave



The information gathered from Figure 3 is presented differently, but also given by Figure 4, which provides a visual for trends in average thermometer levels across waves for each treatment and study combination. Here, it is even more clear that there is a general increase in positive feelings regarding gay marriage on average over time. Although the gap is narrow, the change is evident. To account for possible differences between individuals, we provide a similar graph for a randomly chosen individual in each study-treatment combination group (Figure 5). Having written a function to obtain this graph (see Code Appendix) allows for similar analyses of random trends in each group.

Fig.4: Average Trends in Thermometer Level by Study and Treatment Assignment

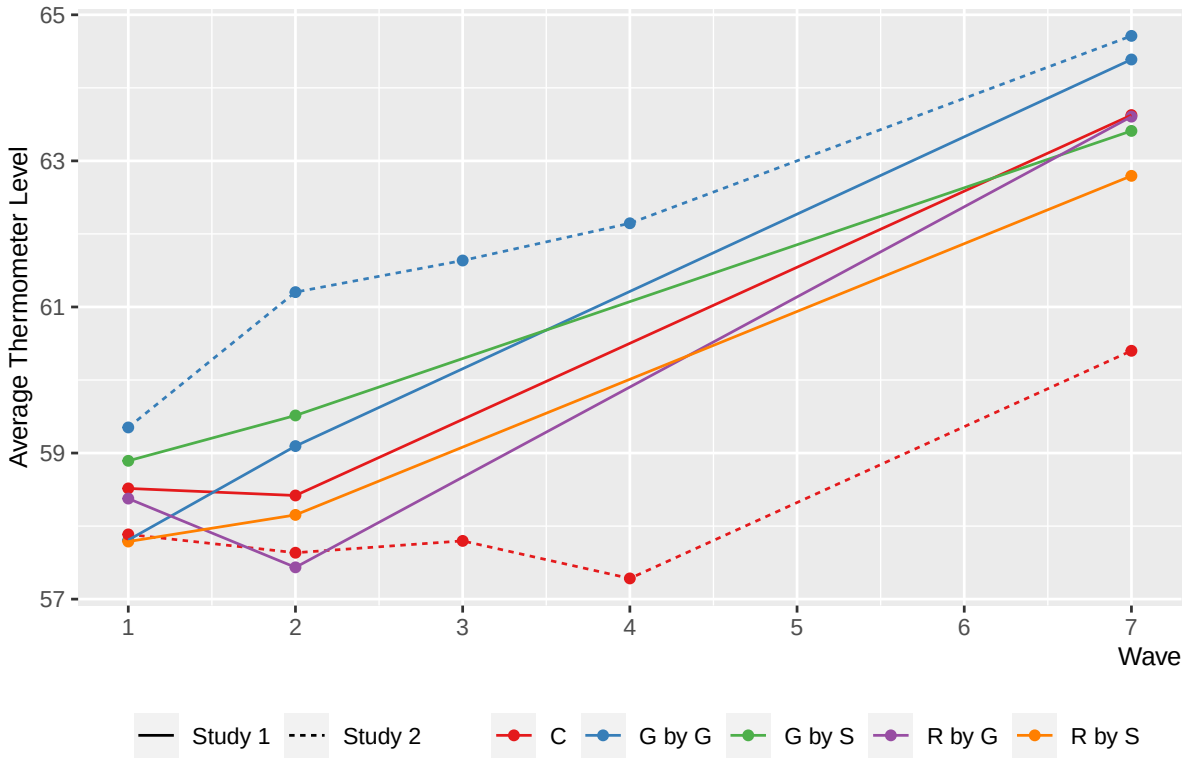
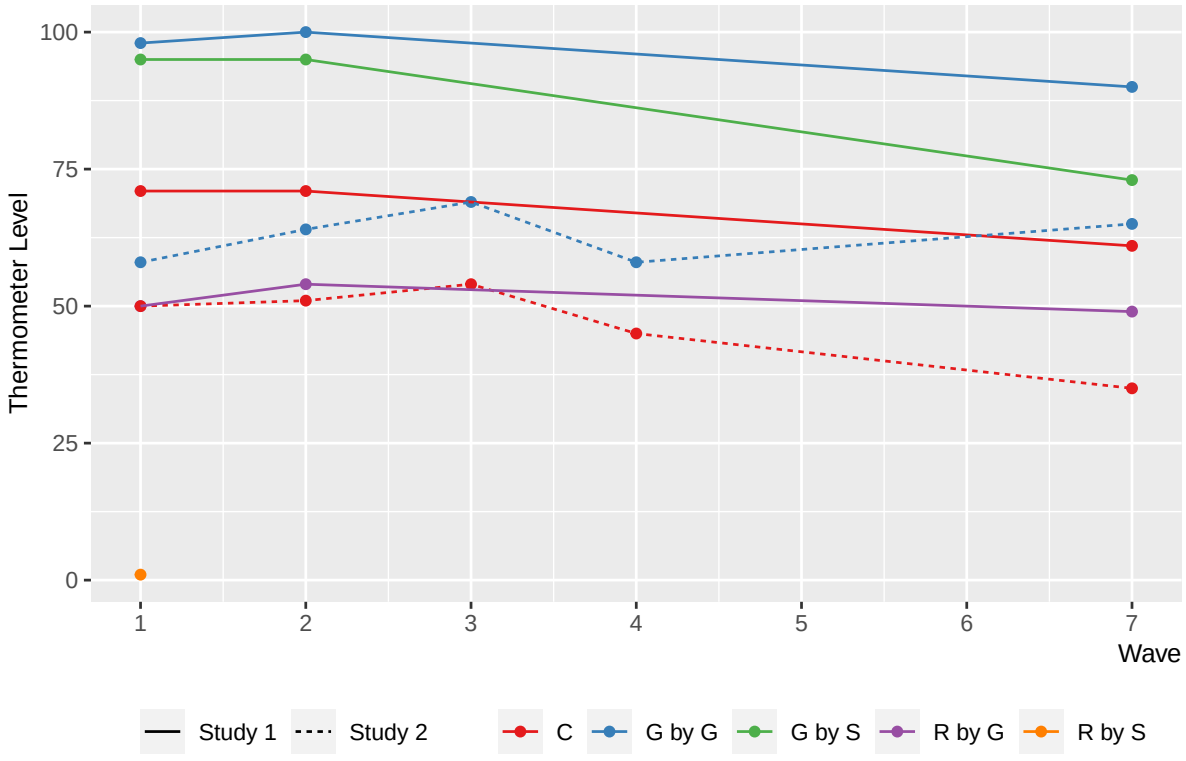


Fig.5: Random Individual Trends in Thermometer Level



Code Appendix

```
## Importing Data

ssm <- read.csv("/Users/antonellabasso/Desktop/PHP2550/Data/ssm_survey.csv")

ssm$panelist_id <- as.factor(ssm$panelist_id) # factorizing "panelist_id"
ssm$wave <- as.factor(ssm$wave) # factorizing "wave"

head(ssm)
```

```
## Descriptive/Summary Statistics

dim(ssm) # 69592 obs, 6 vars
length(unique(ssm$panelist_id)) # 11948 participants

#CreateTableOne(data=ssm) (relevant before factorizing "panelist_id")
#CreateTableOne(data=ssm, strata=c("study"))

# missing values
apply(ssm, 2, function(x) sum(is.na(x))) # 34675 in "therm_level" & "therm_change"
sum(row.names(ssm[is.na(ssm$therm_level),])==
      row.names(ssm[is.na(ssm$therm_change),])) # same rows
# removing NA values as they correspond to unobserved responses
# and do not help with EDA since we are not imputing/bootstrapping
ssm_observed <- ssm[-as.numeric(row.names(ssm[is.na(ssm$therm_level),])),]

# number of observations per participant
obs_id <- as.data.frame(table(ssm_observed$panelist_id)) %>%
  rename("panelist_id"=Var1, "observations"=Freq)
ssm_observed <- full_join(ssm_observed, obs_id) # introducing new var "observations"

# number of participants per observation count (1-5)
num_obs <- as.data.frame(table(obs_id$observations)) %>%
  rename("Observations"=Var1, "Participants ($n$)"=Freq)

# "them_level" & "them_change" descriptive statistics by "study" & "treatment_assignment"
study_ta_desc_stats <- ssm_observed %>%
  group_by(study, treatment_assignment) %>%
  summarise(count=n(),
            mean_tl=round(mean(therm_level), 4),
            sd_tl=round(sd(therm_level), 4),
            mean_tc=round(mean(therm_change), 4),
            sd_tc=round(sd(therm_change), 4))
colnames(study_ta_desc_stats) <- c("Study", "Treatment",
                                   "Participants ($n$)",
                                   "Mean ($\\mu$)",
                                   "SD ($\\sigma$)",
                                   "Mean ($\\mu$)",
                                   "SD ($\\sigma$)")

# "them_level" descriptive statistics by "study", "treatment_assignment", and "wave"
study_ta_wave_desc_stats <- ssm_observed %>%
  group_by(study, treatment_assignment, wave) %>%
```

```
  summarise(count=n(),
            mean=mean(therm_level),
            sd=sd(therm_level))
colnames(study_ta_wave_desc_stats) <- c("Study", "Treatment Assignment", "Wave",
                                       "Participants",
                                       "Average Thermometer Level",
                                       "Thermometer Level Standard Deviation")
```

```
## Distribution Plots

# density plots of "them_level" by "study"
ssm_observed_wide <- pivot_wider(ssm_observed,
                                 names_from="study",
                                 values_from="therm_level")
fig1 <- ggplot(ssm_observed_wide, aes(x=x)) +
  geom_density(aes(x=`Study 1`, y=..density..), fill="#69b3a2", alpha=0.87) +
  geom_text(aes(label="Study 1", x=9, y=0.014, fontface="bold"),
            stat="unique", size=4, color="#69b3a2") +
  geom_density(aes(x=`Study 2`, y=-..density..), fill= "#404080", alpha=0.87) +
  geom_text(aes(label="Study 2", x=9, y=-0.014, fontface="bold"),
            stat="unique", size=4, color="#404080") +
  labs(x="Thermometer Level",
       y="Density",
       subtitle="Fig.1: Thermometer Level Densities by Study") +
       #caption="Fig.1: Thermometer Level Densities by Study") +
  theme(axis.title.x=element_text(size=10.47, hjust=1),
        axis.title.y=element_text(size=10.47))

# histograms of "them_level" by "treatment_assignment"
fig2 <- ggplot(ssm_observed, aes(x=therm_level,
                          color=treatment_assignment,
                          fill=treatment_assignment)) +
  geom_histogram(alpha=0.6, stat="count") +
  scale_fill_viridis(discrete=TRUE) +
  scale_color_viridis(discrete=TRUE) +
  labs(x="Thermometer Level",
       title="Fig.2: Frequency Distributions of TL and TC by Treatment Assignment") +
  theme(axis.title.x=element_text(size=10.47, hjust=1),
        axis.title.y=element_blank(),
        plot.title=element_text(size=11.47),
        legend.position="none",
        panel.spacing=unit(0.1, "lines"),
        strip.text.x=element_text(size=8)) +
  facet_wrap(~treatment_assignment)

# histograms of "therm_change" by "treatment_assignment"
fig2_1 <- ggplot(ssm_observed, aes(x=therm_change,
                          color=treatment_assignment,
                          fill=treatment_assignment)) +
  geom_histogram(alpha=0.6, stat="count") +
  scale_fill_viridis(discrete=TRUE) +
  scale_color_viridis(discrete=TRUE) +
  labs(x="Thermometer Change", title=" ") +
  theme(axis.title.x=element_text(size=10.47, hjust=1),
```

```
        axis.title.y=element_blank(),
        legend.position="none",
        panel.spacing=unit(0.1, "lines"),
        strip.text.x=element_text(size=8)) +
  facet_wrap(~treatment_assignment)

# violin plots of "them_level" density by "study", "treatment_assignment", and "wave"
fig3 <- ggplot(ssm_observed, aes(x=wave,
                         y=therm_level,
                         color=study,
                         fill=study)) +
  geom_violin(alpha=0.74) +
  scale_color_manual(values=c("#69b3a2", "#404080"), aesthetics=c("color", "fill")) +
  labs(x="Wave",
       y="Thermometer Level",
    subtitle="Fig.3: Spread of Thermometer Level by Study, Treatment Assignment, and Wave") +
  theme(axis.title.x=element_text(size=10.47, hjust=0),
        axis.title.y=element_text(size=10.47),
        legend.position=c(0.84, 0.26),
        legend.title=element_blank(),
        panel.spacing=unit(0.1, "lines"),
        strip.text.x=element_text(size=8)) +
  facet_wrap(~treatment_assignment)
```

```
## Trend Line Plots

# "them_level" for each combination of study and treatment assignment across waves

# AVERAGE trend lines
study_ta_wave_avg <- data.frame(study=study_ta_wave_desc_stats$Study,
                                treatment_assignment=
                                  study_ta_wave_desc_stats$`Treatment Assignment`,
                                wave=as.numeric(study_ta_wave_desc_stats$Wave),
                                avg_therm_level=
                                  study_ta_wave_desc_stats$`Average Thermometer Level`)
fig4 <- ggplot(study_ta_wave_avg) +
  geom_point(aes(x=wave, y=avg_therm_level, color=treatment_assignment)) +
  geom_line(aes(x=wave, y=avg_therm_level, color=treatment_assignment, lty=study)) +
  #scale_color_viridis(discrete=TRUE) +
  scale_color_brewer(palette="Set1") +
  scale_x_continuous(breaks=c(1:7)) +
  labs(x="Wave",
       y="Average Thermometer Level",
    subtitle="Fig.4: Average Trends in Thermometer Level by Study and Treatment Assignment") +
  theme(axis.title.x=element_text(size=10.47, hjust=1),
        axis.title.y=element_text(size=10.47),
        legend.title=element_blank(),
        legend.position="bottom")

# INDIVIDUAL trend lines
# selecting one individual at random from each "study" and "treatment_assignment" combination
# FUNCTION
rand_trends <- function(seed, subt){
  set.seed(seed)
```

```
  comb_rand_panelists <- ssm_observed[1,]
  for (i in unique(study_ta_wave_avg$study)){
    for (j in unique(study_ta_wave_avg[study_ta_wave_avg$study==i,
                                       "treatment_assignment"])){
      comb_subset <- ssm_observed[ssm_observed$study==i &
                                    ssm_observed$treatment_assignment==j,]
      panelists <- unique(comb_subset$panelist_id)
      rand_panelist <- sample(panelists, 1)
      rand_panelist_df <- ssm_observed[ssm_observed$panelist_id==rand_panelist,]
      comb_rand_panelists <- rbind(comb_rand_panelists, rand_panelist_df)
    }
  }
  comb_rand_panelists$wave <- as.numeric(comb_rand_panelists$wave)
  comb_rand_panelists <- comb_rand_panelists[-1,]

  ggplot(comb_rand_panelists) +
    geom_point(aes(x=wave, y=therm_level, color=treatment_assignment)) +
    geom_line(aes(x=wave, y=therm_level, color=treatment_assignment, lty=study)) +
    scale_color_brewer(palette="Set1") +
    scale_x_continuous(breaks=c(1:7)) +
    labs(x="Wave",
         y="Thermometer Level",
         subtitle=subt) +
    theme(axis.title.x=element_text(size=10.47, hjust=1),
          axis.title.y=element_text(size=10.47),
          legend.title=element_blank(),
          legend.position="bottom")
}
```

```
#rand_trends(4, "Fig.5.1")
#rand_trends(47, "Fig.5.2")
#rand_trends(72, "Fig.5.3")
#rand_trends(93, "Fig.5.4")

rand_trends(47, "Fig.5: Random Individual Trends in Thermometer Level")
```