Voter Experiment Data Analysis

Prepared by C. Peng & L. Pyott @ WCU

Draft Version: 2023-08-23

Contents

1 Practical Questions

The primary questions to address are:

- Are the voters who applied to vote by mail before any contact randomly distributed among the ten contact groups, as we would expect?
- Do any of the contact methods, alone or in two-way combinations, alter (increase or decrease) the likelihood that someone will apply to vote by mail?
- Do any of the contact methods, alone or in two-way combinations, alter (increase or decrease) the likelihood that someone who applied to vote by mail will vote (either by mail or at the polls)?

The secondary questions of interest • Are the voters who applied to vote by mail before any contact demographically similar to the target population (gender, age, precinct)?

• Are the voters who were not contacted by any method demographically similar to the target population (gender, age, precinct)?

The information to answer these questions isn't in your workbook, but I can provide summary statistics for each of the three groups. I'm just not sure how to test for significant differences, especially as to age and precinct distributions.

2 Creation of Analytic Data

The original `VanID` was removed from the data to create a de-identified data set. The data is uploaded to the GitHub repository.

2.1 Naming Conventions

The final analytic data set is created based on several tabs in the original Excel spreadsheet. We use the following naming conventions to name variables in the final analytic data set.

- Any variable with the suffix `.pp` is from *per protocol table*.
- Any variable with the suffix `.itt` is from *intent to treat table*.
- The suffix `.apply` is for variables that are meant to designate if the voter applied to VBM before or after the treatment.
- The suffix `.voted` is how the voter voted.
- The dependent variable is applied:
- B = applied to VBM before treatment
- A = applied to VBM after treatment
- N = did not apply to VBM

We decided to keep all records in the original data regardless of whether receiving any treatment.

Voters' demographic information is only available for the portion of the voters (5786) from Zone 3 and Zone 6. Since demographic information was aggregated from several distinct tables, a few voters appeared in different tables with different recorded demographic information. We keep only one row from all with duplicated IDs, Therefore the resulting data sets still have 13875 records.

Since about 41.7% voters have demographic information, the demographic information for the rest of the voters was treated as missing values.

```
{ echo = FALSE, eval=FALSE} lp = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester
University of PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\\VoterDataAnalysis\\LP23822.csv")
itt = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester University of PA\\Desktop\\cpeng\\WCU-Teachi
VoteData0 = merge(itt, lp, by = "VanID") ## some demographics of a small portion of
voters demographics01 = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester University
of PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\\VoterDataAnalysis\\demographics01.csv")
demographics02 = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester University of
PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\\VoterDataAnalysis\\demographics02.csv")
demographics03 = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester University of
PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\\VoterDataAnalysis\\demographics03.csv")
demographics04 = read.csv("C:\\Users\\75CPENG\\OneDrive - West Chester University of
PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\\VoterDataAnalysis\\demographics04.csv")
demographics = unique(rbind(demographics01, demographics02, demographics03, demographics04))
## left join VoteData01= merge(VoteData0, demographics, by = "VanID", all.x=TRUE) ##
Keep unique ID VoteData = data_unique(VoteData, select = "VanID") ## write.csv(VoteData,
"C:\\Users\\75CPENG\\OneDrive - West Chester University of PA\\Desktop\\cpeng\\WCU-Teaching\\2023Summer\
## This data was uploaded to the GitHub repository after the VanID was removed
```

2.2 Derived Variables

We define several new variables in this section to check whether each voter received the intended treatment.

- Any variable with the suffix `.okay` is the indicator of whether a voter received the intended treatment.

Since variable `group` in the ITT table is not useful since more than half of voters did not receive the planned treatment(s). We defined the actual treatment group (`pp.group`) based on the intended treatment patterns using the PP table.

`notrt` was defined to reflect whether a voter received treatment. (`1 = no treatment`, '0 = at least one treatment)

```
vot = read.csv("https://raw.githubusercontent.com/pengdsci/VoteExpAnaly/main/DataSet/VoteData.csv")
##
vot$itt.okay = ifelse((vot$mail.itt== vot$mail.pp & vot$phone.itt==vot$phone.pp & vot$lit.itt == vot$li
## Checking compliance
vot$mail.okay = ifelse((vot$mail.itt== vot$mail.pp), "Y", "N")
vot$phone.okay = ifelse((vot$phone.itt==vot$phone.pp), "Y", "N")
vot$lit.okay = ifelse((vot$lit.itt == vot$lit.pp), "Y", "N")
vot$text.okay = ifelse((vot$text.itt==vot$text.pp), "Y", "N")
## The ITT group may not be the same as the actual group in the PP table!
## The analysis should be based on the actual treatment pattern in the PP table
## The next we define `pp.group`
vot$pp.group = ifelse((vot$mail.pp == "Y" & vot$phone.pp == "N" & vot$lit.pp == "N" & vot$text.pp == "N
                    ifelse((vot$mail.pp == "N" & vot$phone.pp == "Y" & vot$lit.pp == "N" & vot$text.pp =
                    ifelse((vot$mail.pp == "N" & vot$phone.pp == "N" & vot$lit.pp == "Y" & vot$text.pp =
                    ifelse((vot$mail.pp == "Y" & vot$phone.pp == "Y" & vot$lit.pp == "N" & vot$text.pp =
                    ifelse((vot$mail.pp == "N" & vot$phone.pp == "N" & vot$lit.pp == "Y" & vot$text.pp =
                    100))))))))))
vot$noTrt = ifelse((vot$mail.pp == "N" & vot$phone.pp == "N" & vot$lit.pp == "N" & vot$text.pp == "N"),
## summarizing
complianceAll = table(vot$itt.okay)
complianceM = table(vot$mail.okay)
complianceP = table(vot$phone.okay)
complianceL = table(vot$lit.okay)
complianceT= table(vot$text.okay)
## comparison between the two `group` variables
complianceGroup = as.matrix(table(vot$group, vot$pp.group))
colnames(complianceGroup) = paste("pp.", c(1:10,100), sep = "")
rownames(complianceGroup) = paste("itt.", 1:10, sep = "")
col.sums = apply(complianceGroup, 2, sum)
row.sums = apply(complianceGroup, 1, sum)
comp.group = rbind(cbind(complianceGroup, RowTotal = row.sums), ColTotal = c(col.sums, sum(col.sums)))
##
list(complianceAll=complianceAll, complianceM=complianceM, complianceP =complianceP, complianceL=complia
```

```
## $complianceAll
##
##    N    Y
## 9182 4693
##
## $complianceM
##
##     Y
## 13875
##
## $complianceP
##
##    N    Y
## 4114 9761
```

```
##
## $complianceL
##
##     N     Y
##  3073 10802
##
## $complianceT
##
##     N     Y
##  3777 10098
##
## $complianceGroup
##          pp.1 pp.2 pp.3 pp.4 pp.5 pp.6 pp.7 pp.8 pp.9 pp.10 pp.100 RowTotal
## itt.1    1459    0    0    0    0    0    0    0    0     0      0     1459
## itt.2       0  346    0    0    0    0    0    0    0     0   1078     1424
## itt.3       0    0  652    0    0    0    0    0    0     0    778     1430
## itt.4       0    0    0  429    0    0    0    0    0     0    968     1397
## itt.5     953    0    0    0  316    0    0    0    0     0      0     1269
## itt.6     729    0    0    0    0  626    0    0    0     0      0     1355
## itt.7     983    0    0    0    0    0  423    0    0     0      0     1406
## itt.8       0  221  479    0    0    0    0  169    0     0    594     1463
## itt.9       0  211    0  327    0    0    0    0  111     0    683     1332
## itt.10      0    0  427  246    0    0    0    0    0   162    505     1340
## ColTotal 4124  778 1558 1002  316  626  423  169  111   162   4606    13875
```

- **A confusion about no-treatment-group**

```
table(vot$noTrt, vot$applied)
```

```
##
##        A    B    N
##   0  213  635 8421
##   1   70  278 4258
```

Why did “before” and “after” treatment indicators appear in the no-treatment group?

- 4606 voters did not receive any treatment at all. We can use the indicator variable `noTrt` to stratify the data and only focus on those who received at least one treatment.

3 Primary Questions

All three questions are based on the `treatment group`. Because the compliance rate is about 47%. The actual treatment group should be defined based on the PP table and used in the analysis.

All variables are categorical, we end up with s few chi-square tests to answer this set of questions.

3.1 Q1

Are the voters who applied to vote by mail before any contact randomly distributed among the ten contact groups, as we would expect?

This is equivalent to testing whether the proportion of voters who applied to vote by mail is equal across the treatment group.

If some of the cells in the frequency table are small, a warning message of the test reliability will be generated.

No treatment group was removed!

```r
q1 = as.matrix(table(vot$pp.group, vot$applied))
before.trt.app = q1[,2]
after.trt.app = q1[,1]
no.app = q1[,3]
applied = before.trt.app + after.trt.app + no.app
prop.test(before.trt.app[-11], applied[-11])
```

```
##
##  10-sample test for equality of proportions without continuity
##  correction
##
## data:  before.trt.app[-11] out of applied[-11]
## X-squared = 10.518, df = 9, p-value = 0.3102
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3     prop 4     prop 5     prop 6     prop 7
## 0.06256062 0.09125964 0.07381258 0.06886228 0.06962025 0.06709265 0.07092199
##     prop 8     prop 9    prop 10
## 0.07692308 0.06306306 0.04938272
```

3.2 Q2

Do any of the contact methods, alone or in two-way combinations, alter (increase or decrease) the likelihood that someone will apply to vote by mail?

This is equivalent to, among voters who received treatments, testing whether the proportions of voters who applied to vote by mail in the corresponding before and after treatment groups are equal across the treatment group.

Several different models can be used for modeling multinomial response. The one we use here is called **the baseline logit model**. It is used for unordered categorical responses and has the following explicit model expressions.

$$\ln\left(\frac{P[\text{applied} = \text{A}]}{P[\text{applied} = \text{N}]}\right) = \alpha_1 + \sum_{i=2}^{10} \alpha_i \text{group}_i$$

$$\ln\left(\frac{P[\text{applied} = \text{A}]}{P[\text{applied} = \text{B}]}\right) = \beta_1 + \sum_{i=2}^{10} \beta_i \text{group}_i$$

R library **nnet** has a function `multinom` the implement this model. We will use this model to report the inferential results. The baseline is chosen to be `N` - not applied to vote by mail.

```r
vot.trt = vot[which(vot$noTrt==0),]
vot.trt$applied.grp <- ifelse(vot.trt$applied =="B", "grp1B",
                              ifelse(vot.trt$applied =="A", "grp2A", "grp3N"))
multilogit =  multinom(applied.grp ~ factor(pp.group), data = vot.trt)
```

```
## # weights:  33 (20 variable)
## initial  value 10183.037304
## iter  10 value 5634.058512
## iter  20 value 4243.944637
## iter  30 value 3322.555919
## iter  40 value 3302.233415
## final  value 3302.231454
```

```
## converged
```

```r
coeff = summary(multilogit)$coefficients
sderr = summary(multilogit)$standard.errors
TStst = coeff/sderr
## Two-tailed normal test for regression coefficients
pval= (1 - pnorm(abs(TStst), 0, 1))*2
model.results = data.frame(
coef.A = t(coeff)[,1],
stder.A = t(sderr)[,1],
TS.A = t(TStst)[,1],
pval.A = t(pval)[,1],
###
coef.N = t(coeff)[,2],
stder.N = t(sderr)[,2],
TS.N = t(TStst)[,2],
pval.N = t(pval)[,2]
)
rownames(model.results) = c("intercept",paste("grp", 2:10, sep = ""))
pander(model.results)
```

Table 1: Table continues below

	coef.A	stder.A	TS.A	pval.A	coef.N	stder.N
intercept	-0.8254	0.1128	-7.318	2.514e-13	2.677	0.06436
grp2	-0.5465	0.2869	-1.905	0.05682	-0.4048	0.1403
grp3	-0.7843	0.2547	-3.079	0.00208	-0.1639	0.1164
grp4	-0.519	0.2877	-1.804	0.07126	-0.09269	0.1404
grp5	-0.06627	0.4114	-0.1611	0.872	-0.1145	0.2306
grp6	-0.1407	0.3148	-0.4469	0.6549	-0.07302	0.1724
grp7	-0.4966	0.4135	-1.201	0.2298	-0.1257	0.2001
grp8	-1.044	0.7678	-1.36	0.174	-0.203	0.2962
grp9	0.266	0.6363	0.4181	0.6759	-0.01981	0.3959
grp10	-0.561	0.7984	-0.7026	0.4823	0.2666	0.3683

	TS.N	pval.N
intercept	41.6	0
grp2	-2.886	0.003904
grp3	-1.409	0.1589
grp4	-0.66	0.5093
grp5	-0.4966	0.6195
grp6	-0.4237	0.6718
grp7	-0.6278	0.5301
grp8	-0.6855	0.493
grp9	-0.05002	0.9601
grp10	0.724	0.4691

3.3 Q3

Do any of the contact methods, alone or in two-way combinations, alter (increase or decrease) the likelihood that someone who applied to vote by mail will vote (either by mail or at the polls)?

We use two different methods to address this question.

- 10-sample before-after comparison: right-tail test of hypothesis

`after.trt.vote-by-mail.rate > before.trt.vote-by-mail.rate`

```
q1 = table(vot$voted, vot$pp.group, vot$applied)
# After treat ment
After.trt = as.matrix(q1[,,1])
a.coltot = apply(After.trt, 2, sum)
#before treatment
Before.trt = as.matrix(q1[,,2])
b.coltot = apply(Before.trt, 2, sum)
a.trt.m.vot = After.trt[1,]/a.coltot
b.trt.m.vot = Before.trt[1,]/b.coltot
## two-sample test for proportions
p = (After.trt[1,] + Before.trt[1,])/(a.coltot+b.coltot)
TS = (a.trt.m.vot - b.trt.m.vot)/sqrt(p*(1-p)/(a.coltot+b.coltot))
## right-tailed test; after treatment rate > before treatment rate
pval = 1-pnorm(TS)
pander(round(cbind(a.trt.m.vot=a.trt.m.vot, b.trt.m.vot=b.trt.m.vot, pval = pval),4))
```

	a.trt.m.vot	b.trt.m.vot	pval
1	0.6283	0.6318	0.555
2	0.6667	0.6197	0.1796
3	0.4348	0.5913	0.9999
4	0.7222	0.6667	0.1337
5	0.6667	0.7727	0.9114
6	0.4375	0.7381	1
7	0.75	0.8667	0.9757
8	0.5	0.6923	0.9429
9	0.75	0.7143	0.3951
10	0.5	0.5	0.5
100	0.6286	0.6367	0.6235

- Binary Logistic Regression

We can also fit a regular logistic regression model to the restricted data that has only two categories. The model formula is

$$\ln\left(\frac{P[\text{applied} = \text{A}]}{P[\text{applied} = \text{B}]}\right) = \gamma_1 + \sum_{i=2}^{10} \gamma_i \text{group}_i$$

```
vot.trt = vot[which(vot$noTrt==0),]
vot.trt.voted = vot.trt[which(vot.trt$applied %in% c("A","B")),]
vot.trt.voted$applied.grp <- ifelse(vot.trt.voted$applied =="B", 0, 1)
logit.voted =  glm(applied.grp ~ factor(pp.group), family = binomial, data = vot.trt.voted)
coeff.voted = summary(logit.voted)$coefficients
pander(coeff.voted)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.8256	0.1128	-7.318	2.509e-13
factor(pp.group)2	-0.5467	0.287	-1.905	0.05677
factor(pp.group)3	-0.7839	0.2548	-3.077	0.002091
factor(pp.group)4	-0.5182	0.2877	-1.801	0.0717

	Estimate	Std. Error	z value	Pr(>|z|)
factor(pp.group)5	-0.06825	0.4114	-0.1659	0.8683
factor(pp.group)6	-0.1395	0.3147	-0.4433	0.6575
factor(pp.group)7	-0.4962	0.4136	-1.2	0.2303
factor(pp.group)8	-1.046	0.7679	-1.362	0.173
factor(pp.group)9	0.266	0.6369	0.4176	0.6762
factor(pp.group)10	-0.5607	0.7986	-0.7022	0.4826

4 Secondary Questions

This analysis is only based on Zone 3 and Zone 6. It seems that the **target population** in the questions is not clearly defined. The following results might not be those that Bob really wanted to see. We can revise this section once we get clarification of **target population**.

```
z36 = which(vot$zone ==3 | vot$zone == 6)
secondayQ = vot[z36,]
```

4.1 Q1

Are the voters who applied to vote by mail before any contact demographically similar to the target population (gender, age, precinct)?

- Gender

```
app.gender = table(vot$sex, vot$applied)
#after.trt.gender = app.gender[,1]
before.trt.gender = app.gender[,2]
applied.gender = app.gender[,1] + app.gender[,2]
prop.test(before.trt.gender, applied.gender )
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  before.trt.gender out of applied.gender
## X-squared = 0.13129, df = 1, p-value = 0.7171
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.07265865  0.04652673
## sample estimates:
##    prop 1    prop 2
## 0.7680891 0.7811550
```

- precinct

```
app.precinct = table(vot$precinct, vot$applied)
#after.trt.gender = app.gender[,1]
before.trt.precinct = app.precinct[,2]
applied.precinct = app.precinct[,1] + app.precinct[,2]
prop.test(before.trt.precinct, applied.precinct )
```

```
## Warning in prop.test(before.trt.precinct, applied.precinct): Chi-squared
## approximation may be incorrect
```

```
##
##  29-sample test for equality of proportions without continuity
```

```
##  correction
##
## data:  before.trt.precinct out of applied.precinct
## X-squared = 33.116, df = 28, p-value = 0.2315
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5    prop 6    prop 7    prop 8
## 0.8846154 0.7230769 0.8372093 0.7692308 0.7391304 0.8266667 0.8666667 0.8571429
##    prop 9   prop 10   prop 11   prop 12   prop 13   prop 14   prop 15   prop 16
## 0.7037037 0.7410714 0.8333333 0.3333333 0.7826087 0.6930693 0.6666667 0.6500000
##   prop 17   prop 18   prop 19   prop 20   prop 21   prop 22   prop 23   prop 24
## 0.7521368 0.7777778 0.7666667 0.6363636 0.8666667 0.8333333 0.9000000 0.5000000
##   prop 25   prop 26   prop 27   prop 28   prop 29
## 0.7846154 0.8139535 0.7192982 0.7714286 0.8076923
```

- Age

The **target population** is not clearly defined. We discretize the age in the following;

```
vot$grp.age = cut(vot$age, c(18,28,38,48,58,68,78,10000), labels = NULL,
    include.lowest = TRUE, right = TRUE)
A.id = which(vot$applied == "A")
B.id = which(vot$applied == "B")
id = which(vot$applied =="N")
App.dat = vot[-id,]
app.age = as.matrix(table(App.dat$applied, App.dat$grp.age))
age.tot = apply(app.age, 2, sum)
b.app = app.age[2,]
prop.test(b.app, age.tot)
```

```
##
##  7-sample test for equality of proportions without continuity correction
##
## data:  b.app out of age.tot
## X-squared = 16.523, df = 6, p-value = 0.01121
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4    prop 5    prop 6    prop 7
## 0.6202532 0.7012987 0.8148148 0.7592593 0.7955556 0.8109453 0.8041237
```

```
#hist(vot$age)
```

4.2 Q2

Are the voters who were not contacted by any method demographically similar to the target population (gender, age, precinct)? – This quetion seems to be ill-posed!.

```
no.contact.dat = vot[id,]
```

- Gender

```
table(no.contact.dat$sex)/sum(table(no.contact.dat$sex))
```

```
##
##         F         M
## 0.6269789 0.3730211
```

- grp.age

```
table(no.contact.dat$grp.age)/sum(table(no.contact.dat$grp.age))
```

```
##
##    [18,28]    (28,38]    (38,48]    (48,58]    (58,68]    (68,78] (78,1e+04]
## 0.10191293 0.15369393 0.19129288 0.20184697 0.18436675 0.12104222 0.04584433
```

- precinct

```
per = 100*round(table(no.contact.dat$precinct)/sum(table(no.contact.dat$precinct)),4)
precinct.name = names(per)
percentage =paste(per,"%", sep="")
pander(data.frame(precinct = precinct.name, Percentage = percentage) )
```

precinct	Percentage
Birmingham 1	2.52%
Birmingham 2	3.64%
East Nottingham East	4.93%
East Nottingham West	4.36%
Elk	1.96%
Franklin	6.45%
Highland	1.22%
London Britain	5.1%
London Grove Ch	2.35%
London Grove S	8.86%
Londonderry	2.85%
Lower Oxford E	1.62%
Lower Oxford W	2.17%
New London	7.56%
Oxford E	2.05%
Oxford W	4.4%
Penn	7.33%
Thornbury 1	1.49%
Thornbury 2	2.96%
Upper Oxford	2.61%
W Fallowfield	1.89%
W Nottingham	2.18%
West Grove 1	2.09%
West Grove 2	1.55%
Westtown 1	3.15%
Westtown 2	3.68%
Westtown 3	4.71%
Westtown 4	2.48%
Westtown 5	1.85%