Clinic Sales Experiment Analysis

Shih-Yuan Wang

A. Brief Introduction of Clinic Sales Experiment

Einfach Medical Supplies (EMS) sells a disposable testing device to clinics and is trying to expand their business in a new region. They want to find a better way to pitch the product when visiting a clinic. A Clinic Sales Experiment was designed to test whether different opening pitches have different effects on conversion of clinics by using a randomized controlled design. So in this new region the company took the clinics and randomized them into three groups:

a) Control (value 0) that received the original "cost-focused" pitch;
b) Treatment 1 (value 1) that received the "easier to use" pitch;
c) Treatment 2 (value 2) that received the "fewer errors" pitch.

When the sales team went out to visit a clinic, they saw the pitch that was randomly selected for that clinic on their work iPad.

To simplify, the three main parts of the experiment were as follows:

a) Target population: clinics in this new region
b) Treatment: received the "easier to use" pitch (1) or the "fewer errors" pitch (2)
c) Outcome metrics: conversion rate (percentage of clinics that did purchase test kits from EMS)

```r
# Load libraries
library(knitr)
library(dplyr)
library(skimr)
library(jtools)
library(rcompanion)
library(ggplot2)
library(car)
```

```r
rm(list = ls()) # clear the workspace

x <- paste("C:/Users/User/Desktop/BUS 740 - Experiments and Causal Methods for Business Insights/",
           "Assignment/Homework 1_Clinic Sales Case Study", sep="")
setwd(x)
```

B. Experiment Analysis

1. Read in the file and clean the data.

```
clinicsales <- read.csv('clinicsales.csv', header = TRUE)  # load the data file
kable(summary(clinicsales))  # give a descriptive summary for all variables in clinicsales
```

	clinicnumber	treatment	purchase	numdoctors	avgpanelsize	distance
	Min. : 1.00	Min. :0.000	Min. :0.0000	Min. :1.000	Min. :2050	Min. : 2.00
	1st Qu.: 81.25	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2186	1st Qu.: 9.00
	Median :161.50	Median :1.000	Median :0.0000	Median :4.000	Median :2308	Median : 21.00
	Mean :161.50	Mean :1.016	Mean :0.2764	Mean :3.935	Mean :2304	Mean : 35.31
	3rd Qu.:241.75	3rd Qu.:2.000	3rd Qu.:0.7500	3rd Qu.:5.000	3rd Qu.:2420	3rd Qu.: 45.75
	Max. :322.00	Max. :2.000	Max. :9.0000	Max. :7.000	Max. :2546	Max. :437.00

```
# Exclude the row where the purchase variable is not 0 or 1 (data error)
clinicsales <- clinicsales[clinicsales$purchase %in% c(0,1), ]
```

There are 6 variables and 322 observations in the initial dataset. When reviewing a few quick summary statistics, I found that the purchase variable contains erroneous value 9, and I removed this row. After cleaning the data, there are 321 remaining observations in the dataset. (Please see the following summary table.)

2. Create a table of summary statistics for the variables in the data.

```
# summary statistics showing the mean and standard deviation of each variable
skim(clinicsales)
```

Table 2: Data summary

Name	clinicsales
Number of rows	321
Number of columns	6
Column type frequency:	
numeric	6
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
clinicnumber	0	1	161.73	93.15	1	81	162	242	322	
treatment	0	1	1.01	0.82	0	0	1	2	2	
purchase	0	1	0.25	0.43	0	0	0	0	1	
numdoctors	0	1	3.93	1.17	1	3	4	5	7	
avgpanelsize	0	1	2303.27	141.64	2050	2186	2308	2419	2546	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
distance	0	1	35.38	43.08	2	9	21	46	437	

From the summary statistics, we can see that around 25% of the clinics did purchase test kits from EMS, and the average number of doctors working at the clinic was around 4. The mean of the average number of patients each of the doctors at the clinic has in their panel was 2303, and the average distance between the clinic and the sales office was around 35.4 miles.

3. Create a table of the shares of clinics in each of the experimental treatment/control groups.

```
attach(clinicsales) # attach the dataset

tb_treatment <- matrix(NA, nrow = 2, ncol = 3) # create a empty output matrix with 2 rows
# (for Frequency, i.e., count, and Proportion) and the 3 groups
tb_treatment[1,] <- formatC(round(table(treatment)), format="f", digits=0) # counts in treatment.
tb_treatment[2,] <- round(prop.table(table(treatment)), 3) # proportion in treatments
rownames(tb_treatment) <- c("Frequency", "Proportion" ) # name the rows
colnames(tb_treatment) <- c("Cost-focused (Control)", "Easier to use", "Fewer errors")
# name the columns
kable(tb_treatment, align = "rrr") # output the table in a readable format
```

	Cost-focused (Control)	Easier to use	Fewer errors
Frequency	106	105	110
Proportion	0.33	0.327	0.343

There are roughly 33% of clinics were randomized into each of the treatments, for around 107 clinics in each of the treatment and control groups.

4. Check for balance in the pre-experiment variables across treatment groups.

For a successful experiment, the pre-treatment variables should be balanced across treatment and control groups. The variables of interest here are number of doctors (numdoctors), average number of panel size (avgpanelsize), and distance. We are hoping to have these look similar between the groups, since they can't be affected by the treatment.

a. Create a table comparing the means between treatment and controls.

```
# create a data frame with the treatment variables and the pre-treatment variables
preexp <- clinicsales %>%
  select(numdoctors, avgpanelsize, distance) # only use these variables for summary in clinicsales

# Summarize the means of those variables by treatment
tb_preexp <- matrix(NA, nrow = 3, ncol = 3) # define the empty output matrix
colnames(tb_preexp) <- c("Mean Cost-focused (Control)", "Mean Easier to use", "Mean Fewer errors")
# name the columns
rownames(tb_preexp) <- colnames(preexp) # name the rows
```

```
m <- as.matrix(round(aggregate(.~treatment, preexp, mean), 2))
# summarize all variables by treatment and store the outputs as a matrix

tb_preexp[,1:3] <-t(m)[2:4,]  # transpose the matrix and delete the treatment row
kable(tb_preexp) # output the table in a readable format
```

	Mean Cost-focused (Control)	Mean Easier to use	Mean Fewer errors
numdoctors	3.75	3.98	4.05
avgpanelsize	2291.79	2282.62	2334.04
distance	38.03	35.16	33.05

We can see that everything looks roughly balanced here. The averages for each of these three variables look similar across each of the groups.

As sometimes averages can hide something important, we plot the histogram for these pre-experiment variables (numdoctors, avgpanelsize, and distance) to check whether the pre-treatment variables are really balanced across treatment and control groups.

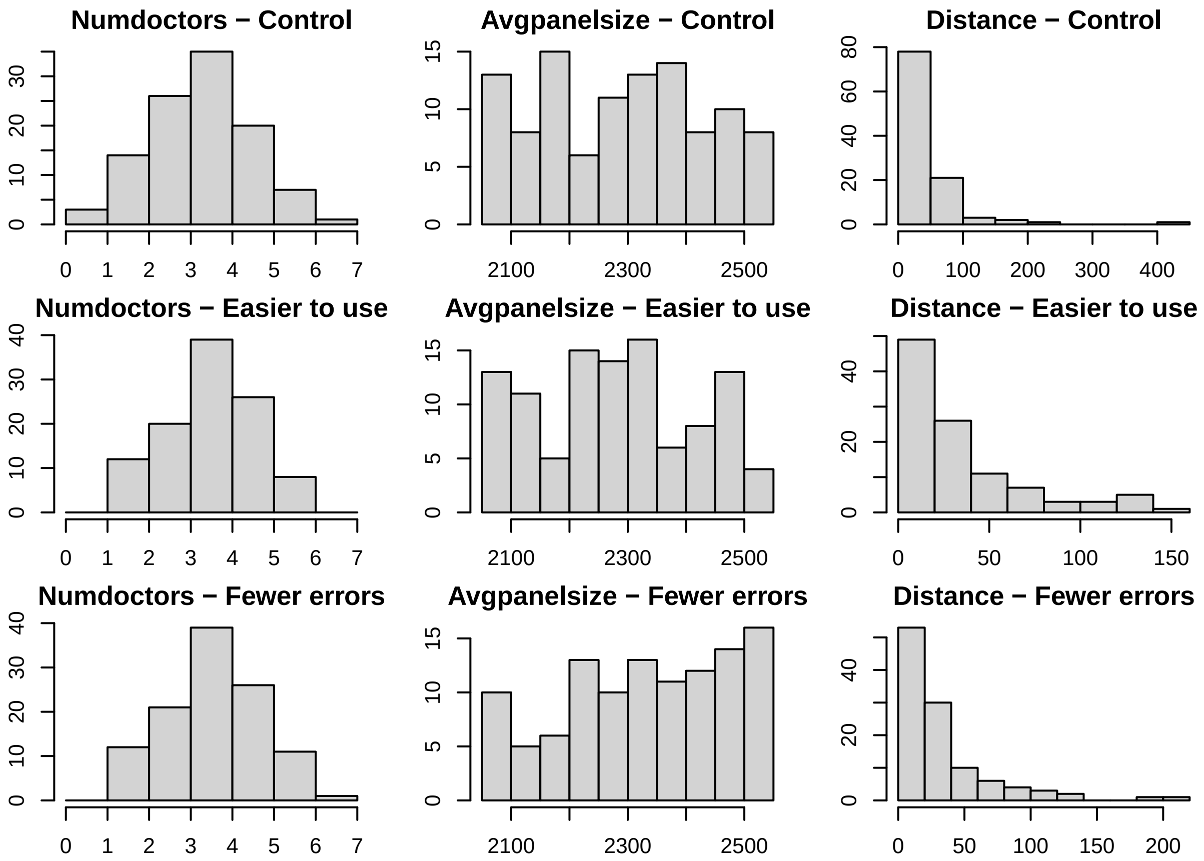
b. Graph the histograms of each variable separately for treatment and control groups.

```
par(mar = c(2,2,2,2))
par(mfrow=c(3,3)) # output multiple subfigures into one figure,
# with 3 subfigures each row and 3 rows (one for each treatment) in total

# plot the histogram of numdoctors for control group
hist(numdoctors[treatment==0], main = paste("Numdoctors - Control"), xlab = "Control",
     breaks = seq(0, 7))
# plot the histogram of avgpanelsize for control group
hist(avgpanelsize[treatment==0], main = paste("Avgpanelsize - Control"), xlab = "Control")
# plot the histogram of distance for control group
hist(distance[treatment==0], main = paste("Distance - Control"), xlab = "Control")

# plot the histogram of numdoctors for treatment 1 group
hist(numdoctors[treatment==1], main = paste("Numdoctors - Easier to use"), xlab = "Easier to use",
     breaks = seq(0, 7))
# plot the histogram of avgpanelsize for treatment 1 group
hist(avgpanelsize[treatment==1], main = paste("Avgpanelsize - Easier to use"), xlab = "Easier to use")
# plot the histogram of distance for treatment 1 group
hist(distance[treatment==1], main = paste("Distance - Easier to use"), xlab = "Easier to use")

# plot the histogram of numdoctors for treatment 2 group
hist(numdoctors[treatment==2], main = paste("Numdoctors - Fewer errors"), xlab = "Fewer errors",
     breaks = seq(0, 7))
# plot the histogram of avgpanelsize for treatment 2 group
hist(avgpanelsize[treatment==2], main = paste("Avgpanelsize - Fewer errors"), xlab = "Fewer errors")
# plot the histogram of distance for treatment 2 group
hist(distance[treatment==2], main = paste("Distance - Fewer errors"), xlab = "Fewer errors")
```



By looking down each of these columns of histograms, we can see that the distributions of the variable numdoctors are roughly similar, though not a very high degree of similarity. However, the distributions of the variable avgpanelsize and distance are not quite similar. This might because of the small sample size. If we could get larger sample size, both the averages and the distributions might have looked more similar and balanced. In the following analysis, we would assume that the experiment was completely randomized and their outcome differences will be due to the "treatment".

5. Plot the means and confidence intervals of "purchased" by control and both treatments.

The main thing we were interested in with this experiment is getting an estimate of how the percentage of clinics that did purchase test kits from EMS (conversion rate) varies with different pitch methods.

We start by creating a summary table that has the means and 95% confidence intervals for the outcome for all of the treatments.

```
# Create a summary table
summary <- clinicsales %>%  # create a table called summary that will hold the info
  mutate(treatment = as.factor(treatment)) %>% # treatment is a factor variable taking discrete levels
  group_by(treatment) %>%                      # create groups by treatment
  summarise(n = length(clinicnumber),                    # create a new table with summary measures
            avgpercpurchased = round(mean(purchase), 3), # get the mean for each group
            error = round(sd(purchase)/sqrt(n), 3),      # calculate the standard error on the mean
  # using the standard deviation divided by square root of n in each group
            lowerCI = round(avgpercpurchased - 1.96*error, 3),  # calculate 95% CI boundaries
            upperCI = round(avgpercpurchased + 1.96*error, 3))
```

```r
kable(summary, caption = "**Average Conversion Rate (Percentage of Purchased)**")
```

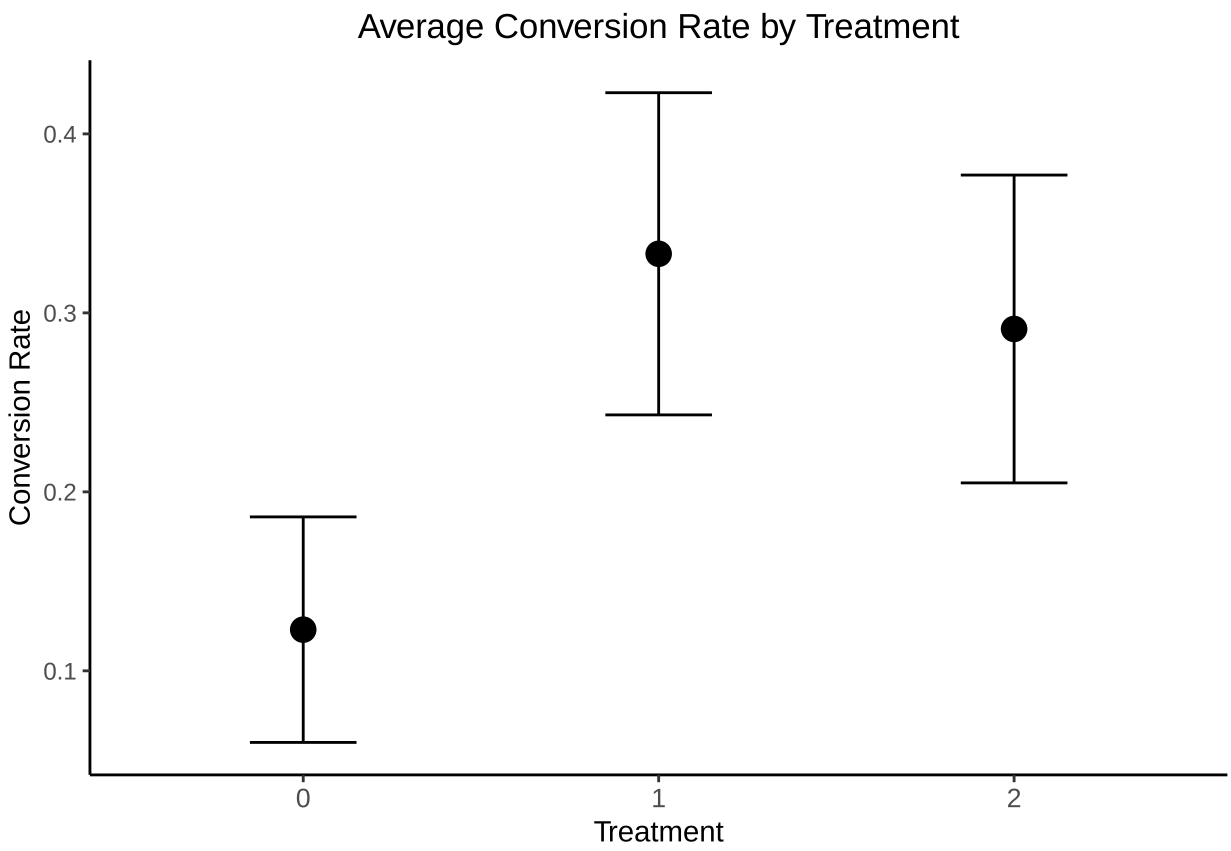
Table 6: **Average Conversion Rate (Percentage of Purchased)**

treatment	n	avgpercpurchased	error	lowerCI	upperCI
0	106	0.123	0.032	0.060	0.186
1	105	0.333	0.046	0.243	0.423
2	110	0.291	0.044	0.205	0.377

So we now have a table with our experimental results. We can see that the average percentage of clinics that did purchase test kits rises with both the “easier to use” (33.3%) and the “fewer errors” (29.1%) treatment groups, compared to the original “cost-focused” (12.3%) control group. Also, the confidence intervals on the averages of the treatment groups do not overlap with those of the control group. It indicates that using “easier to use” or “fewer errors” pitch method would likely lead to higher conversion rate. However, the averages of these two treatment groups are close, and the confidence intervals on the averages of them are wide and overlap, so we are unsure about which treatment is better even though the average of the “easier to use” group is a little bit higher.

Next, we produce a visual of the results by creating a graph of the means with their 95% confidence intervals.

```r
# Plot the means and confidence intervals of "purchased" from that summary table
summary %>%
  ggplot(aes(x = treatment)) +
  geom_point(aes(y = avgpercpurchased), size = 4) +
  scale_shape_manual(values=c(15, 16)) +
  ggtitle("Average Conversion Rate by Treatment") +
  ylab("Conversion Rate") + xlab("Treatment") +
  theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"),
        axis.text.x = element_text(size = 10), legend.position=c(.5,.5),
        plot.title = element_text(hjust=.5)) +
  geom_errorbar(aes(ymin = lowerCI,
                    ymax = upperCI), width = .3) +
  scale_color_manual(values=c("darkgrey", "black"))
```



6. Calculate the Average Treatment Effect (ATE) for each treatment relative to control and provide a 95% confidence intervals on the ATE.

The estimate of the Average Treatment Effect (ATE) is the difference between the average for each treatment group and the average for the control group. The 95% confidence intervals on the ATE means that the interval has 95% chance of containing the true difference in the expected conversion rate.

```
ATE <- matrix(NA, ncol = 3, nrow = 3)  # create a matrix to store the results
colnames(ATE) <- c("Treatment Effect", "Lower 95% CI", "Upper 95% CI" )
rownames(ATE) <- c("Easier to use", "Fewer errors", "Control Mean")

# calculate the average treatment effect
effect <- c(summary$avgpercpurchased[2]-summary$avgpercpurchased[1],
            summary$avgpercpurchased[3]-summary$avgpercpurchased[1],
            summary$avgpercpurchased[1])
error_ate <- c(sqrt(summary$error[1]^2+summary$error[2]^2),
               sqrt(summary$error[1]^2+summary$error[3]^2), NA)

# calculate the standard error of ATE
LCI <- effect - 1.96*error_ate
UCI <- effect + 1.96*error_ate

ATE[,1] <- round(effect,3)
ATE[,2] <- round(LCI,3)
ATE[,3] <- round(UCI,3)
```

```
kable(ATE, caption = "**Average Treatment Effect on Conversion Rate**" )
```

Table 7: **Average Treatment Effect on Conversion Rate**

	Treatment Effect	Lower 95% CI	Upper 95% CI
Easier to use	0.210	0.100	0.320
Fewer errors	0.168	0.061	0.275
Control Mean	0.123	NA	NA

We can find that relative to the original "cost-focused" pitch, the "easier to use" treatment increases conversion rate by 0.21, and the "fewer errors" treatment increases conversion rate by 0.168. For the "easier to use" treatment group, the interval (0.1, 0.32) has 95% chance of containing the true difference in the expected conversion rate. On the other hand, for the "fewer errors" treatment group, the interval (0.061, 0.275) has 95% chance of containing the true difference in the expected conversion rate. Besides, both treatments have a statistically positive significant effect on conversion of clinics at the 95% confidence level.

C. Conclusion

To sum up, the experiment results imply that we are 95% sure that either the "easier to use" pitch (treatment 1) or the "fewer errors" pitch (treatment 2) would have higher conversion of clinics than the original "cost-focused" pitch (control) even though we are still uncertain about which treatment is better than the other. We can also estimate that the expected conversion rate of the "easier to use" pitch is around 33.3% and the expected conversion rate of the "fewer errors" pitch is around 29.1%, while the expected conversion rate of the original "cost-focused" pitch is just around 12.3%.

D. Notes for Caveats and Limitations

The distributions of the pre-treatment variable "avgpanelsize" and "distance" are not quite similar among the control and treatment groups. As the randomly selected groups did not have good balance of each variable, the differences in their conversion rate (percentage of purchased) might be due to the other confounders including the average panel size and distance, which would bias the conclusions from the experiment. If we could get larger sample size, both the averages and the distributions might have looked more similar and balanced. In this experiment, we assume that the experiment was completely randomized and their outcome differences will be due to the "treatment".