Clinic Sales Experiment Analysis

Annie Chen

2023-10-23

Overview

This document presents the analysis of data from Einfach Medical Supplies (EMS), which aimed to sell test kits to medical clinics in a new region using different sales approaches in an experiment.

The business problem is that we need to test whether different sales approaches have different effects on conversion of clinics. Initially, the company used the "cost saving" approach for selling test kits. Now, we are considering whether to change the methods into emphasizing "easier to use" or "fewer errors" in the sales pitch.

In this experiment, sales team were randomly assigned to different pitching methods for clinics:

- **Control**: Clinics received the original "cost-focused" pitch.
- **Treatment 1**: Clinics received the "easier to use" pitch.
- **Treatment 2**: Clinics received the "fewer errors" pitch.

By conducting this experiment, the company aims to effectively convey the product pitch to doctors and nurse practitioners in the clinic, with the ultimate goal of boosting test kit sales and increasing the clinic conversion rate.

The first step is to load the data:

```r
clinic_data = read.csv("/Users/anniechen/Desktop/clinicsales.csv", header = TRUE)
```

Looks like it loaded!

Calculate summary statistics

We can use summarize:

```r
summary = summary(clinic_data)

kable(summary, digits = 2)
```

	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

	clinicnumber	treatment	purchase	numdoctors	avgpanelsize	distance
	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

For the column "purchase", it takes value 0 if the clinic declined to purchase test kits and 1 if they did purchase from EMS. However, the maximum in column "purchase" is 9 as shown above. We should address this concerning value by removing incorrect data or outliers:

```
clinic_data <- clinic_data[clinic_data$purchase %in% c(0, 1), ]
```

In this way, we keep only the rows where "purchase" is 0 or 1.

Let's summarize the data again:

```
summary = summary(clinic_data)
num = nrow(clinic_data)

kable(summary, digits = 2)
```

	clinicnumber	treatment	purchase	numdoctors	avgpanelsize	distance
	Min. : 1.0	Min. :0.000	Min. :0.0000	Min. :1.000	Min. :2050	Min. : 2.00
	1st Qu.: 81.0	1st Qu.:0.000	1st Qu.:0.0000	1st Qu.:3.000	1st Qu.:2186	1st Qu.: 9.00
	Median :162.0	Median :1.000	Median :0.0000	Median :4.000	Median :2308	Median : 21.00
	Mean :161.7	Mean :1.012	Mean :0.2492	Mean :3.931	Mean :2303	Mean : 35.38
	3rd Qu.:242.0	3rd Qu.:2.000	3rd Qu.:0.0000	3rd Qu.:5.000	3rd Qu.:2419	3rd Qu.: 46.00
	Max. :322.0	Max. :2.000	Max. :1.0000	Max. :7.000	Max. :2546	Max. :437.00

```
print(num)
```

```
## [1] 321
```

There are a total of 321 clinics in this data set. The treatment takes value 0 for control, 1 for Treatment 1, 2 for Treatment 2. There are an average 3.931 doctors per clinic. Average Panel size is around 2303 patients each of the doctors has in their panel. Distance ranges from 2 to 437 miles.

Look for Balance

```
  balance_table = clinic_data %>% group_by(treatment) %>% summarise(
    mean_doc = mean(numdoctors),
    mean_panelsize = mean(avgpanelsize),
    mean_dis = mean(distance),
    N = n(),
    prop = n()/322
    )

kable(balance_table, digits = 2)
```

treatment	mean_doc	mean_panelsize	mean_dis	N	prop
0	3.75	2291.79	38.03	106	0.33
1	3.98	2282.62	35.16	105	0.33
2	4.05	2334.04	33.05	110	0.34

The data look pretty balanced across treatment arms. Doctor averages very close to 4 in every arm. Panel size close to 2300, and distance close to 35. We can also see the proportion in each arm are very close to 33 percent, which is quite balanced!

Calculate Outcome Mean and Confidence Interval in each arm

The formula for the 95% Confidence Interval is: $\overline{x} \pm 1.96 s_x/\sqrt{N}$.

```
arm_info = clinic_data %>% group_by(treatment) %>% summarise(
  mean_p = mean(purchase),
  sd_p = sd(purchase),
  N = n(),
  lb_p = mean_p - 1.96 * sd_p / sqrt(N),
  ub_p = mean_p + 1.96 * sd_p / sqrt(N)
  )
meanCI = arm_info %>% select(treatment, mean_p, lb_p, ub_p)

kable(meanCI, digits = 2)
```

treatment	mean_p	lb_p	ub_p
0	0.12	0.06	0.19
1	0.33	0.24	0.42
2	0.29	0.21	0.38

Notably, treatment 1 appears to have a higher mean compared to other treatments and control. There is evidence that treatment 1 has a statistically significant impact on clinics purchasing decision.

Calculate ATEs and CIs

The formula for the 95% Confidence Interval is: $\overline{x} \pm 1.96 s_x/\sqrt{N}$.

Moreover, the formula for standard error of ATE is: $\sqrt{\frac{sd_t^2}{N_t} + \frac{sd_c^2}{N_c}}$

```
ate_info = arm_info %>% mutate(
  ate = mean_p - mean_p[1],
  se_ate = sqrt( sd_p^2 / N + sd_p[1]^2 / N[1]),
  lb = ate - 1.96 * se_ate,
  ub = ate + 1.96 * se_ate
  )
ate_info = ate_info %>% filter(treatment != 0) %>% select(treatment, ate, lb, ub)

kable(ate_info, digits = 2)
```

treatment	ate	lb	ub
1	0.21	0.10	0.32
2	0.17	0.06	0.27

We can see the treatment 1 ("easier to use" approach) stands out with a higher mean purchase rate compared to the control group and treatment 2. The average treatment effect (ATE) for treatment 1 is 0.21 with a relatively narrow confidence interval. That is, treatment 1 is statistically significant because the confidence interval for this effect is relatively tight, suggesting a high level of precision in the estimate. This suggests that the company should use the "easier to use" pitch which might resulted in a significant increase in clinic conversions, with a high degree of confidence.

Caveats and Limitations

- **Confounding Variables**: The analysis focused on the impact of different pitching approaches on clinic conversions. However, there might be other confounding or intermediate variables that is not included in the data set could influence the outcomes. For example, different types of clinics may prefer to use different kinds of test kits. Moreover, the economic conditions in the region might also be the confounding variable that lead clinics to be more budget-conscious and hesitant to make purchases, regardless of the sales approach.
- **Overall Evaluation Criteria (OEC)**: The analysis focused on the outcome of whether the clinic purchase the test kits or not. However, there might be some long-term effects or combination of outcomes that is sufficient for decision making. The lasting effects or potential changes in clinic behavior are unexplored in this experiment.
- **Proxy Outcomes**: The analysis should consider proxy outcome that is not itself directly relevant but highly correlated with the desired outcome. For example, customers' feedback, clinic overall revenue, and new patient rates could be the proxy outcomes that provide a more comprehensive understanding of the impact of different sales approaches.