Prices of Personal Computers

Introduction

For our study, the population parameter of interest we are finding is the difference between the proportions of all computers with CD-ROM and all computers without CD-ROM that have a multimedia kit.

$$p_{CD-ROM} - p_{Non\ CD-ROM}$$

For our point estimate in our samples, we are finding the difference between the sampled proportions of computers with a CD-ROM and for those that do not have a CD-ROM that have a multimedia kit.

$$\hat{p}_{CD-ROM} - \hat{p}_{Non\ CD-ROM}$$

The research question for this study is, “is there a difference in the proportion of computers with a CD-ROM with a multi-media kit and computers without a CD-ROM that have a multi-media kit?”

The null hypothesis is, “computers with a CD-ROM have a multimedia kit and computers without CD-ROM that have a multimedia kit have the same proportion.”

$$H_0:\ p_{CD\ ROM} = p_{non\ CD\ ROM}$$

The alternative hypothesis is, “computers with a CD-ROM have a multimedia kit and computers without CD-ROM that have a multimedia kit have a difference of proportion.”

$$H_A:\ p_{CD\ \&\ Multi} \neq p_{non\ CD\ \&\ Multi}$$

The Sample Data

The dataset we are working with includes a total number of 6,259 computers. There are 10 variables such as the price, speed (clock speed in MHz), hd (hard drive size in MB), ram (size of ram in MB), screen (screen size in in.), cd (has a CD-ROM), multi (has a multimedia kit), premium (has premium), ads (number of 486 price listings each month), and trend (time trend from Jan. 1993 to Nov. 1995). This data was gathered to analyze intertemporal pricing and price discrimination in the computer market.

This data can be collected from: https://vincentarelbundock.github.io/Rdatasets/datasets.html

And for the documentation:

https://vincentarelbundock.github.io/Rdatasets/doc/Ecdat/Computers.html

The variables that we are interested in are cd and multi. They are both categorical variables and have two levels such as “yes” and “no.” Yes as in they are included with the computer and no that they are not. From our research question, we are seeing whether or not there is a difference in the proportions of multimedia kit computers that have a CD-ROM and for those that do not have a CD-ROM.

The Methodology

The method I will be using to test my hypothesis is the difference of two proportions and I will test the conditions that will meet a normal distribution. For the confidence level, we will subtract 5% from 100% of the whole area under the curve of the normal distribution to get 95%. We then divide 5% by 2 to get 2.5% for each of the two tails.

The formulas we will use are the confidence interval formula and the standard error formula which finds how many standard errors the data is away from the true value which is 0.

$$CI = pt.\ est. \pm z^* * SE$$

$$pt.\ est. = (\hat{p}_{CD} - \hat{p}_{Non\ CD})$$

$$SE = \sqrt{\frac{\hat{p}_{CD}(1-\hat{p}_{CD})}{n_{CD}} + \frac{\hat{p}_{Non\ CD}(1-\hat{p}_{Non\ CD})}{n_{Non\ CD}}}$$

The Condition

The conditions that must be met with a difference of two proportions using a confidence interval is by assuming that both computers with CD ROM and non CD ROM are independent. We may also assume that each group is less than 10% of the whole population size. There must also be at least 10 successes and 10 failures for each group according to the rule of thumb. Most importantly, the sample size has to be big enough in order for the distribution to be nearly normal according to the central limit theorem.

We first grab a subset for each group by using the command below.

```
cd <- subset(Computers, cd == "yes")
non_cd <- subset(Computers, cd == "no")
```

There are a total of 2908 observations for computers with CD ROM and 3351 observations for computers without CD ROM. We create n1 and n2 for the sample size in each group

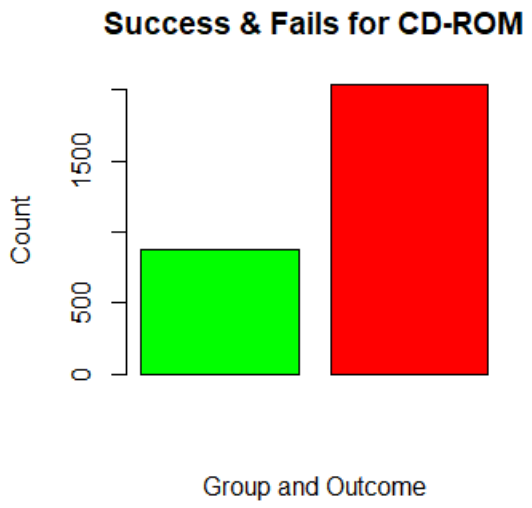
```
n1 <- 2908

n2 <- 3351
```

Below shows the command to find the number of successes and failures of CD-ROM computers that have a multimedia kit.

```
success1 <- sum(cd$multi == "yes")
 fails1 <- sum(cd$multi == "no")
```

There are 873 successes and 2035 failures in the CD-ROM group. Below shows the bar charts side by side of the number of successes and failures.



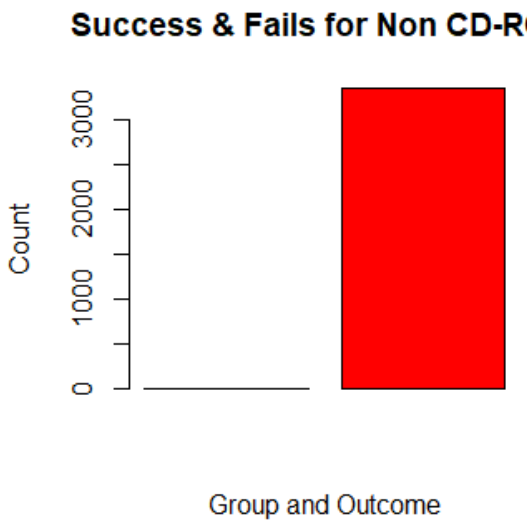

```
counts1 <- c(success1, fails1)

barplot(counts1, beside = TRUE, col = c("green", "red"), main = "Success &
Fails for CD-ROM", xlab = "Group and Outcome", ylab = "Count")
```

Below shows the command to find the number of successes and failures of non CD-ROM computers that have a multimedia kit.

```
success2 <- sum(non_cd$multi == "yes")
  fails2 <- sum(non_cd$multi == "no")
```

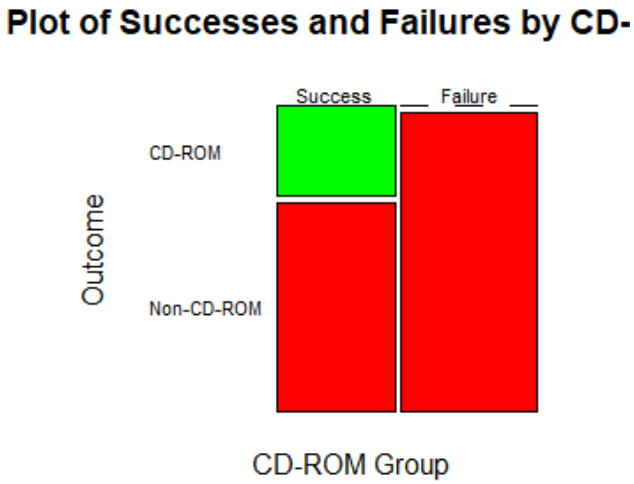
There are 0 successes and 3351 failures in the non CD-ROM group. Below shows the box plots side by side of the number of successes and failures.




```
counts2 <- c(success2, fails2)

barplot(counts2, beside = TRUE, col = c("green", "red"), main = "Success &
Fails for Non CD-ROM", xlab = "Group and Outcome", ylab = "Count")
```

The computers with no CD-ROM do not have at least 10 successes which means the condition is not met. Since the condition is not met, we shouldn't expect the data to be nearly distributed and we may also run to an error.



```
data_matrix <- matrix(c(success1, fails1, success2, fails2), nrow = 2, byrow =
TRUE)
rownames(data_matrix) <- c("Success", "Failure")
colnames(data_matrix) <- c("CD-ROM", "Non-CD-ROM")
mosaicplot(data_matrix, main = "Mosaic Plot of Successes and Failures by
CD-ROM Group", color = c("green", "red"), xlab = "CD-ROM Group", ylab =
"Outcome", las = 1)
```

The mosaic plot above also shows that computers without a CD-ROM do not have a single multimedia kit.

Multimedia	CD-ROM	Non CD-ROM
Multimedia Kits	873	0
No Multimedia Kits	2035	3351
Total	2908	3351
$\widehat{p}$	0.30	0

The table above shows both CD and non CD groups that have multimedia kits which are the successes. The sampled proportions are calculated below.

```
pHat1 <- success1 / n1
pHat2 <- success2 / n2
```

The Confidence Interval

To calculate the confidence interval, we will use the formula below

$$CI = pt.\ est. \pm z^{*} * SE$$

We first find the point estimate by the difference between the proportions of sampled CD-ROM computers and non CD-ROM computers that have a multimedia kit.

$$pt.\ est. = 0.30 - 0 = 0.30$$

```
ptEst <- pHat1 - pHat2
```

To find the critical value using a 95% confidence level, we subtract 95% from 100% to get 5% of the area of one tail, then divide by 2 to get 2.5% for each tail area.

We subtract 2.5% from 100% to get 0.975.

According to the z distribution table, we get a z score of 1.96. We can assign this value to the variable z. I use a z-score because the sample size is large.

```
z <- 1.96
```

For the standard error which calculates how many standard errors the point estimate is away from the true value is calculated below. The larger the standard error, the more spread out the data is. The smaller the standard error means the closer the data is from the true value.

$$SE = \sqrt{\frac{0.30(1-0.30)}{2908} + \frac{0(1-0)}{3351}} = 0.0085$$

```
SE <- sqrt(((pHat1*(1-pHat1))/(n1))+((pHat2*(1-pHat2))/(n2)))
```

Now to calculate the confidence interval that has 95% confidence in the true value of the difference in two proportions of both CD and non CD-ROM groups with a multimedia kit, we input the values in the formula below.

$$CI = 0.30 \pm 1.96 * 0.0085 = (0.283, 0.317)$$

```
lower <- 0.30-1.96*0.0085
upper <- 0.30+1.96*0.0085
```

Here we see that the confidence interval is between 0.283 and 0.317. We are 95% confident that the true value is within the interval. The upper and lower values of the interval are very close together and do not include 0 which means we reject the null hypothesis in favor of the alternative hypothesis. The 0 means there is no difference in both proportions.

This would mean that computers with a CD-ROM have a multimedia kit and computers without CD-ROM that have a multimedia kit have a difference of proportion.
However, since the condition for the rule of thumb is not met, this will lead to an error in our interval.

If we had at least 10 successes in our non CD-ROM group, grabbing many random samples from our population will most likely show a distribution that is nearly normal.

A possible future research is not possible when using this methodology with CD-ROMs and non CD-ROMs and their proportion of multimedia kits due to the failed condition. Even though the standard error of the point estimate is 0.0085 which means how far away its spread from the true value 0, we cannot confirm that this is accurate. Instead, we can use the methodology when comparing CD and non CD-ROMs with the proportion of those that have premium. The rule of thumb is met for that proportion which is guaranteed for the distribution to be nearly normal.