HW4-Darshan Patel-3:30-5:30PM

Darshan Patel

11/1/2018

The company Performance Tires plans to engage in direct mail advertisting. It is currently in negotiations to purchase a mailing list of the names of people who bought sports cars within the last three years. The owner of the mailing list claims that sales generated by contacting names on the list will more than pay for the cost of using the list. (Typically, a company will not sell its list of contacts, but provides the mailing services. For example, the owner of the list would handle addresing and mailing catalogs.)

Before it is willing to pay the asking price of $3 per name, the company obtains a random sample of 225 names and addresses from the list to run a small experiment. It sends promotional mailings to each of these customers. The company makes a profit of 20% on the gross dollar of a sale (not including the $3 cost of the name). For example, if a customer orders $100 worth of goods (i.e., gross dollar), the company makes a $20 profit. If we include the cost of the name, the $20 profit reduces to a $17 profit. Should the company purchase the mailing list?

Question 1:

Why would a company want to run an experiment? Why not just buy the list and see what happens?

Answer: A company would want to run an experiment to assess what would happen if an action would fully be done. In this case, by running an experiment using a small sample of names and addresses, the company can see whether sending promotional mailings would make profit for the company. By doing this, the company spends less money and thus create less loss if the experiment fails, i.e., customers do not order goods. If the company buys the whole list, then the company is at a greater loss if the promotional mailings does not work to its full extent.

Question 2:

Why would the holder of the list agree to allow the potential purchaser to run an experiment?

Answer: The holder of the list would agree to allow the potential purchaser to run an experiment because if the company's experiment is successful, then the company would be likely to come back and purchase the entire list at its asking price.

Question 3:

If you wanted to run a hypothesis test on the profitability of the list at the $\alpha = 0.05$ level, what would your hypotheses be? What does μ represent?

Answer: The null hypothesis is: $H_0: \mu = \$0$ per name. The alternative hypothesis is: $H_A: \mu > \$0$ per name. μ represents the average profit the company makes.

Question 4:

Identify the population, parameter, sample and statistics in this scenario.

Answer: The population is the full list of customer names and addresses the holder has of its clients. The sample is the mailing list of 225 names and addresses from the population. The parameter is $\mu = \$0$, the average profit earned per name and the statistics is the actual average profit earned per name from the sample.

Question 5:

In the hypotheses from question 3, what would it mean to make a Type I error in this context? What is the probability of making such an error?

Answer: A Type I error is created when the null hypothesis is rejected when it is true. In this context, it would mean rejecting the hypothesis that the company makes $0 profit and concluding that the company does make a positive profit when it actually makes $0 profit. The probability of making such an error is 5%.

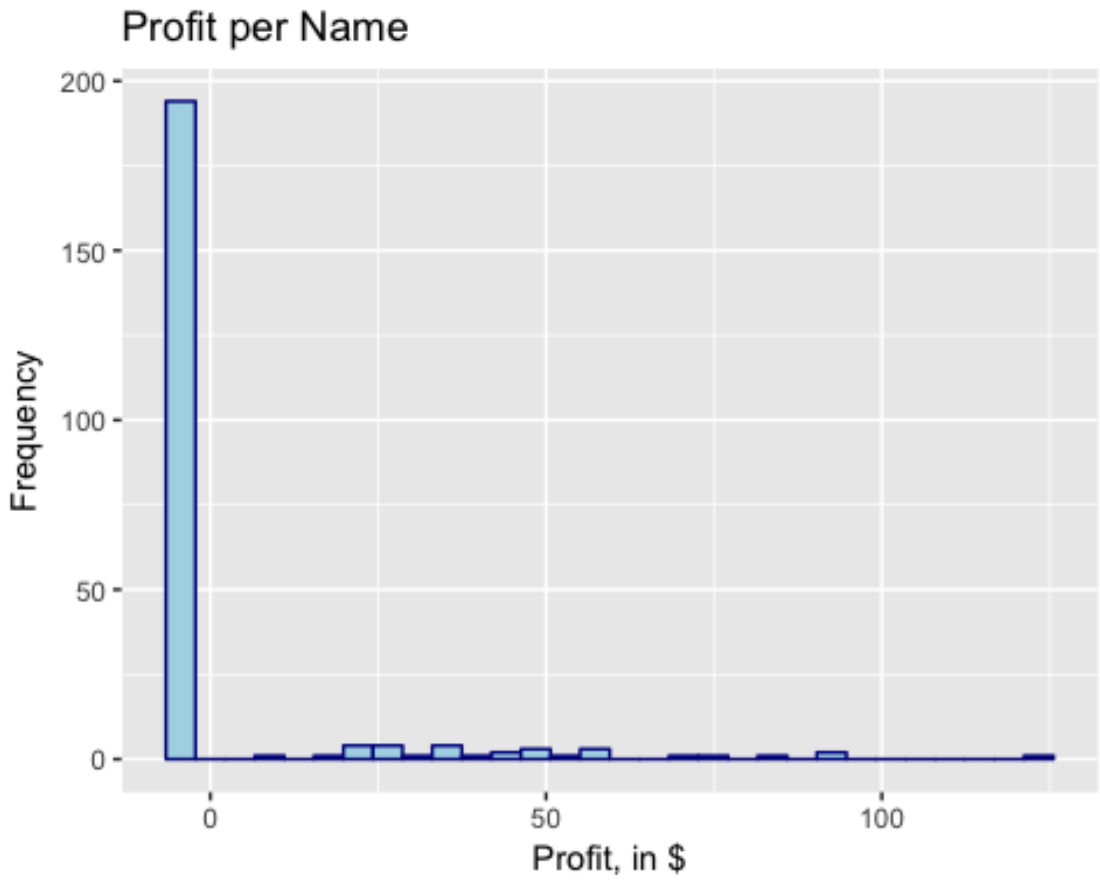
Question 6:

With the data to test the hypotheses, (a) construct a histogram, (b) compute summary statistics (minimum, median, mean, maximum, and standard deviation), and (c) compute the fraction of people who bought nothing from Performance Tires. Describe the shape of the histogram. Include the units of measurement.

Answer:

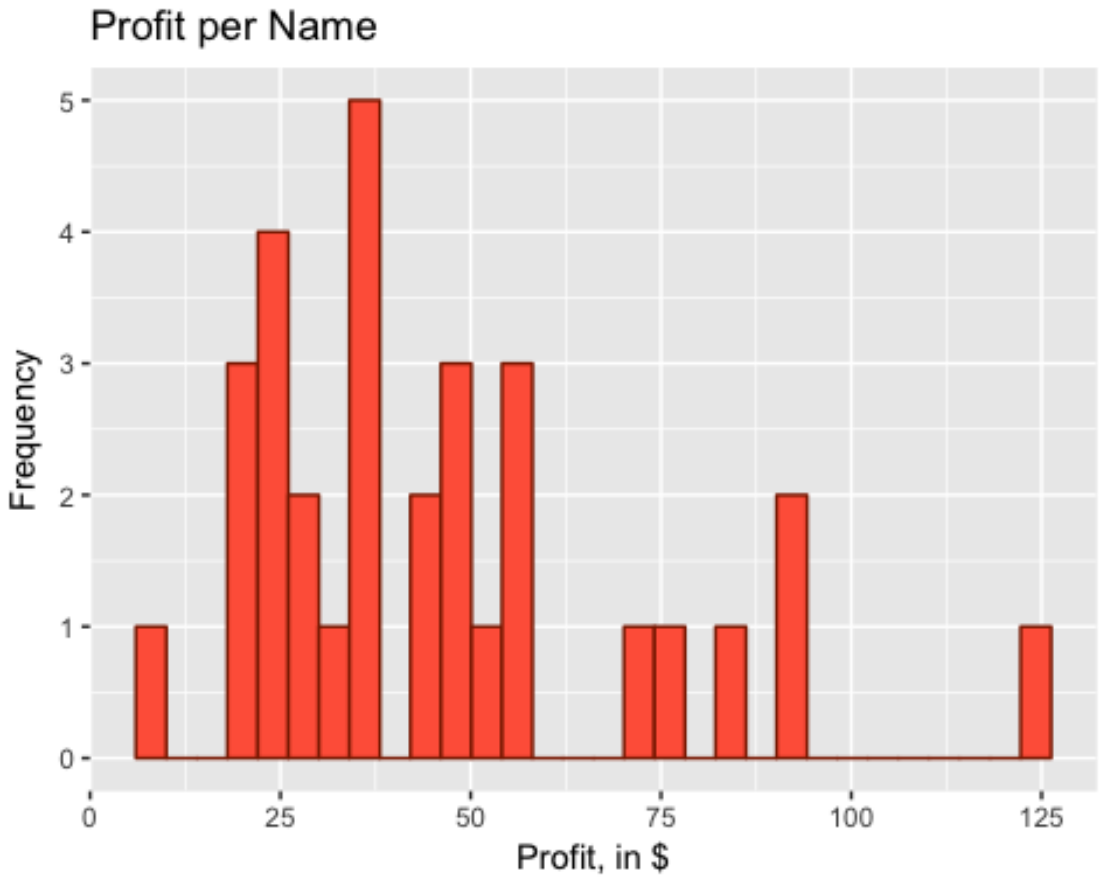
```
# Import packages
library(tidyverse)

## ── Attaching packages
─────────────────────────────────────────────────────────────────────
─────────────────────────────────────────────────────────────────────
──────────────────────── tidyverse 1.2.1 ──

## ✔ggplot2 2.2.1     ✔purrr   0.2.5
## ✔tibble  1.4.2     ✔dplyr   0.7.7
## ✔tidyr   0.8.2     ✔stringr 1.3.0
## ✔readr   1.1.1     ✔forcats 0.3.0

## Warning: package 'tidyr' was built under R version 3.4.4

## Warning: package 'purrr' was built under R version 3.4.4

## Warning: package 'dplyr' was built under R version 3.4.4

## ── Conflicts
─────────────────────────────────────────────────────────────────────
```

```
──────────────────────────────────────────────────────────────────────────────
──────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()

library(readxl)

## Warning: package 'readxl' was built under R version 3.4.4

# Read in the excel file
df <- read_excel("direct_mail.xlsx")

# Create a df for the actual profit the company makes per for each order
profit_per_name <- (df * .2) - 3

# Plot a histogram showing the frequency of profit made in the sample
ggplot(profit_per_name, aes(order_cost)) + geom_histogram(color = 'darkblue',
fill ='lightblue') + ggtitle("Profit per Name") + labs(x = "Profit, in $", y
= "Frequency")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The shape of the histogram of the profit per name is heavily skewed right. To aid the above visual and get a clearer image, let's take out the people who did not make any purchases.

```
# Filter out the people who did not make any order and choose the ones that
did purchase something
purchasers <- data.frame(order_cost = profit_per_name[profit_per_name > 0])

# Plot a histogram showing the frequency of profit made from people who have
purchased a good in the sample
ggplot(purchasers, aes(order_cost)) + geom_histogram(color = 'orangered4',
fill ='tomato1') + ggtitle("Profit per Name") + labs(x = "Profit, in $", y =
"Frequency")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on this subset of the data, the shape of the histogram is skewed right, which agrees with the above statement.

The summary statistics of the profit per name is:

```
# Store the summary statistics of the profit per names in a nice easy to read
dataframe
ss <- data.frame("Minimum" = min(profit_per_name$order_cost),
                 "Median" = median(profit_per_name$order_cost),
                 "Mean" = mean(profit_per_name$order_cost),
                 "Maximum" = max(profit_per_name$order_cost),
                 "Standard Deviation" = sd(profit_per_name$order_cost))
```

```
# Print the transpose of above dataframe for good looking view in pdf
t(ss)
```

```
##                          [,1]
## Minimum             -3.000000
## Median              -3.000000
## Mean                 3.732871
## Maximum            124.784000
## Standard.Deviation  19.403914
```

The fraction of people who brought nothing from Performance Tires is:

```
# Get the percentage of people who brought nothing from Performance Tires
paste(round(100 * length(profit_per_name[profit_per_name <= 0]) /
length(profit_per_name$order_cost), 3), '%', sep = '')
```

```
## [1] "86.222%"
```

Question 7:

Check the assumptions for a one-sample t-test. Are they satisfied for this data? Explain.

Answer: This data comes from a simple random sample. The company has received a list of 225 randomly chosen names and addresses from the owner of the mailing list. In addition, the data is normally distributed. Since $n = 225 > 30$, by the Central Limit Theorem, the sampling distribution is normally distributed.

Question 8:

Test the hypotheses from question 3 and provide a recommendation to the company. Identify the test statistic, degrees of freedom, p-value and conclusion.

Answer:

```
# Individually calculate all variables for calculating the test statistics
for a one-sample T test
xbar <- ss$Mean
mu_naught <- 0
s <- ss$Standard.Deviation
n <- length(profit_per_name$order_cost)

# Calculate the test statistics
t <- (xbar - mu_naught) / (s / sqrt(n))

# Calculate the p-value, or P(T > t), using the T distribution with degrees
of freedom given
deg_free <- n - 1
p_val <- 1 - pt(t, df = n-1)
```

The test statistics is

```
t
```

```
## [1] 2.885658
```

The degrees of freedom is

```
deg_free
```

```
## [1] 224
```

and the p-value is

```
p_val
```

```
## [1] 0.00214381
```

To verify above result, check using `t.test`.

```
# One line code to perform one-sample T test, used to verify above result
t.test(profit_per_name$order_cost, alternative = "greater", mu = 0,
conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  profit_per_name$order_cost
## t = 2.8857, df = 224, p-value = 0.002144
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  1.596261      Inf
## sample estimates:
## mean of x
##  3.732871
```

Conclusion: At the α level of 0.05, because p-value $= 0.002144 < \alpha = 0.05$, the null hypothesis is rejected. Meaning, the company will make more than \$0 profit per name. Therefore, the company should buy the complete mailing list of names and addresses.

Question 9:

What is the probability of making a Type II error with the hypothesis test in question 3 if the average profit was actually \$2?

Answer:

```
# Import package for power
library(pwr)

# Calculate the power of a test for the alternative of profit being $2
power <- pwr.t.test(n, d = (2 - 0)/ss$Standard.Deviation, sig.level = 0.05,
type = "one.sample", alternative = "greater")$power

# Get probability of making a type II error and print it
typeII_error <- 1 - power
paste(round(typeII_error * 100, 3), '%', sep = '')
```

```
## [1] "54.119%"
```

The probability of making a Type II error with the hypothesis test as given when the average profit was actually $2 is around 54.119%.