ECO394D Probability and Statistics Homework 4

Sibo Ding

Summer 2023

Problem 1

Part A

Question: Does one of "Living with Ed" and "My Name is Earl" have a higher mean `Q1_Happy` than the other?
Approach: 2-sample two-sided t-test
Results:

```
##
##  Welch Two Sample t-test
##
## data:  lwe and mni
## t = 1.1676, df = 162.57, p-value = 0.2447
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1030341  0.4011371
## sample estimates:
## mean of x mean of y
##  3.926829  3.777778
```

Conclusion: p-value is greater than 0.05 (or 95% confidence interval includes 0). We cannot reject H_0 so no one show has a higher mean `Q1_Happy` at 5% significance.

Part B

Question: Does one of "The Biggest Loser" and "The Apprentice: Los Angeles" have a higher mean `Q1_Annoyed` than the other?
Approach: 2-sample two-sided t-test
Results:

```
##
##  Welch Two Sample t-test
##
## data:  tbl and tal
## t = -2.1032, df = 300.66, p-value = 0.03628
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.52455614 -0.01743792
## sample estimates:
## mean of x mean of y
##  2.036232  2.307229
```

Conclusion: p-value is smaller than 0.05 (or 95% confidence interval does not include 0). We reject H_0 so one show has a higher mean `Q1_Annoyed` at 5% significance.

Part C

Question: Use a filtered data set to infer the proportion of 4 or grater `Q2_Confusing` of "Dancing with the Stars" with its 95% confidence interval.
Approach 1: 1-sample proportions test
Results:

```
##
##  1-sample proportions test with continuity correction
##
## data:  sum(dwt$Q2_Confusing >= 4) out of length(dwt$Q2_Confusing >=
4), null probability 0.5
## X-squared = 127.65, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.04453431 0.12893254
## sample estimates:
##          p
## 0.07734807
```

Approach 2: Normal approximation for binomial distribution (based on C.L.T.)
Results:
Actual proportion $p =$

```
## [1] 0.07734807
```

Sample size $n =$

```
## [1] 181
```

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) = N(0.0773, 0.0199^2)$$
$$E(\hat{p}) = p = 0.0773$$
$$CI(p)_{0.95} = p \pm z\sqrt{\frac{p(1-p)}{n}} = 0.0773 \pm 1.96 \times 0.0199$$

```
## Lower limit: 0.03842989

## Upper limit: 0.1162662
```

Approach 3: Use bootstrap and Monte Carlo simulation to generate many samples and estimate C.I..

Problem 2

Question: Whether the revenue ratios are the same in the treatment and control groups?
Approach: 2-sample two-sided t-test
Results:

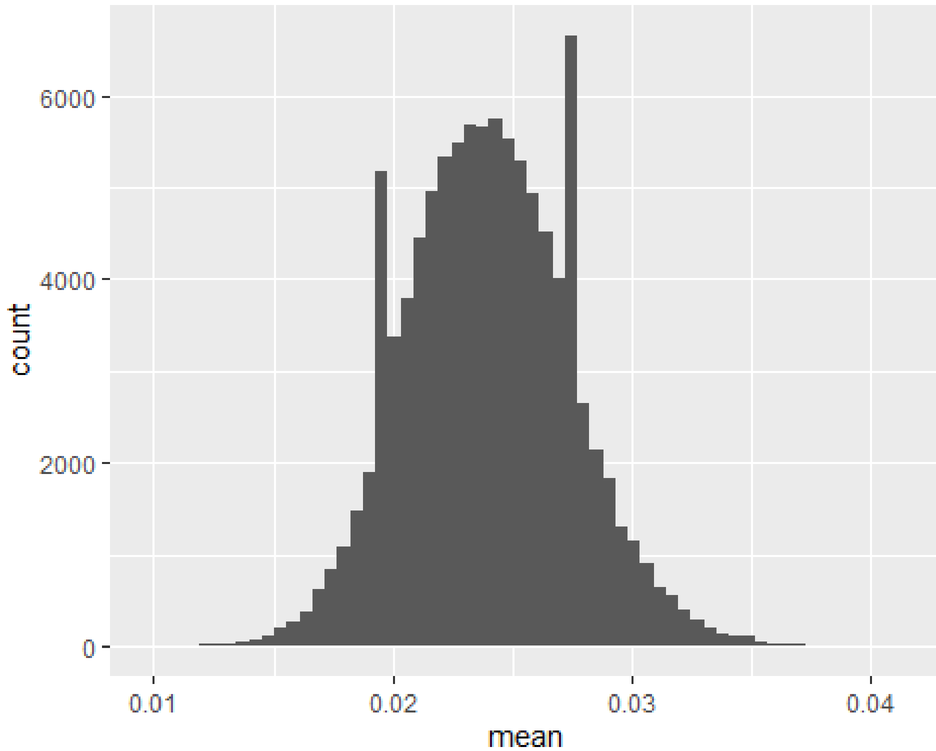
```
##
##  Welch Two Sample t-test
##
## data:  control and treatment
## t = 2.6367, df = 110.2, p-value = 0.00958
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.01298714 0.09157576
## sample estimates:
## mean of x mean of y
## 0.9488775 0.8965961
```

Conclusion: p-value is smaller than 0.05 (or 95% confidence interval does not include 0). We reject H_0 so the revenue ratios in the treatment and control groups are different at 5% significance.

Problem 3

H_0: The proportion of flagged trades from Iron Bank is 2.4%.
Test statistic: $rate = \frac{70}{2021}$
When plotting simulation results, histogram rather than p.d.f. makes more sense:

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
##   fortify.SpatialPolygonsDataFrame ggplot2
```

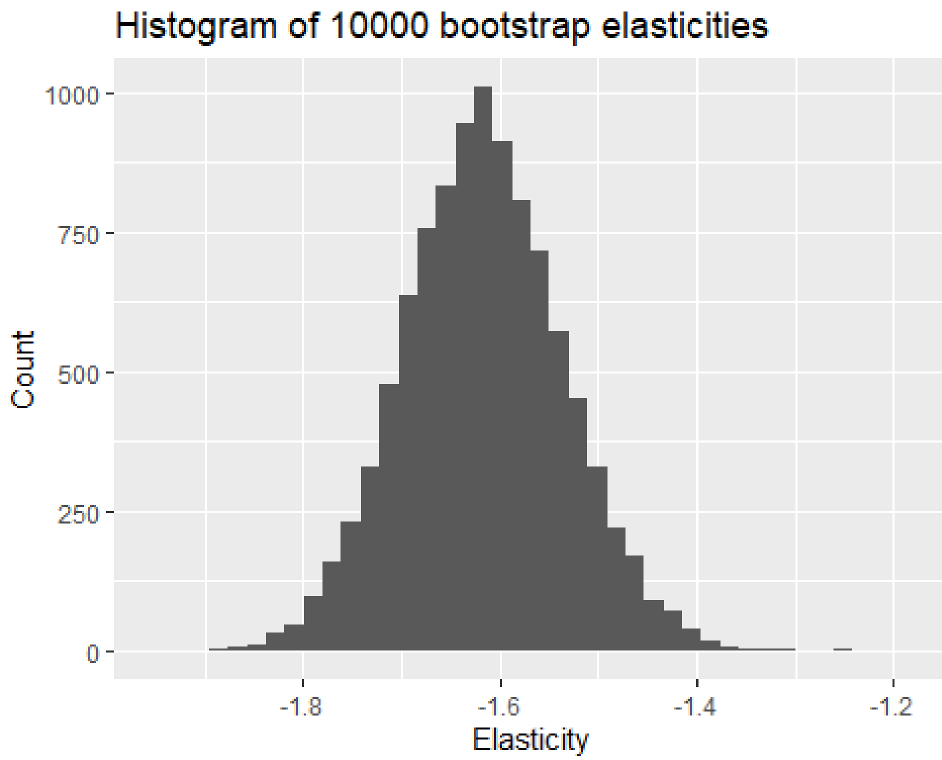


p-value for observing $rate > \frac{70}{2021} =$

```
## [1] 0.00129
```

Conclusion: H_0 is not plausible in light of the data because $p < 0.01$ is pretty small, which is very rare to happen.

Problem 4




```
## Confidence Interval from Bootstrap Distribution (10000 replicates)

##             2.5% 97.5%
## percentile -1.77 -1.45
```

Problem 5

Part A

i. $\because X_1, \ldots, X_N \sim i.i.d. Bernoulli(p)$

$$E(\hat{p}) = E(\bar{X}_N) = E\left(\frac{X_1 + \cdots + X_N}{N}\right)$$
$$= \frac{1}{N}\big(E(X_1) + \cdots + E(X_N)\big)$$
$$= \frac{1}{N} NE(X)$$
$$= p$$

Similarly, $E(\hat{q}) = q$, thus $E(\hat{p} - \hat{q}) = E(\hat{p}) - E(\hat{q}) = p - q$

ii. Since $X_1, \ldots, X_N \sim i.i.d. Bernoulli(p)$ having the same finite mean and variance, as $N \to \infty$, based on C.L.T., $se(\bar{X}_N) = \frac{sd(X)}{\sqrt{N}} = \sqrt{\frac{p(1-p)}{N}}$

iii. $\because X_1, \dots, X_N \sim i.i.d. Bernoulli(p), Y_1, \dots, Y_M \sim i.i.d. Bernoulli(q)$

Based on C.L.T., $Var(\hat{p}) = \frac{Var(X)}{N} = \frac{p(1-p)}{N}, Var(\hat{q}) = \frac{Var(Y)}{M} = \frac{q(1-q)}{M}$

$$\therefore Var(\hat{p} - \hat{q}) = Var(\hat{p}) + Var(\hat{q}) = \frac{p(1-p)}{N} + \frac{q(1-q)}{M}$$

$$se(\hat{p} - \hat{q}) = \sqrt{\frac{p(1-p)}{N} + \frac{q(1-q)}{M}}$$

Part B

Similar to Part A,

$$E(\bar{X}_N - \bar{Y}_M) = E(\bar{X}_N) - E(\bar{Y}_M) = \mu_X - \mu_Y$$

$$Var(\bar{X}_N - \bar{Y}_M) = Var(\bar{X}_N) + Var(\bar{Y}_M) = \frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}$$

$$se(\bar{X}_N - \bar{Y}_M) = \sqrt{\frac{\sigma_X^2}{N} + \frac{\sigma_Y^2}{M}}$$