ST551: HOMEWORK 1

Brian Cervantes Alvarez
October 13, 2023

Question 1

Suppose an observational study showed that people who carry lighters have a higher rate of lung cancer. There are four possibilities:

1. The association is spurious: this was just a particularly unusual dataset.
2. The observed association is causative: carrying lighters causes lung cancer.
3. The causality in the observed association is reversed: lung cancer causes carrying lighters.
4. There is a confounder: something else about people who carry lighters is the true cause of lung cancer. Perhaps people who carry lighters are more likely to be smokers, and so smoking is a confounding variable: it is associated with both the predictor (carrying lighters) and the outcome (lung cancer).

In each of the following settings, suggest a plausible alternative to the causal relationship suggested by the study: either propose a possible confounding factor, or describe why the causality might be reversed.

Part A

Observational studies show that smokers have a higher rate of liver cancer. Suggested causal relationship: Smoking causes liver cancer.

Answer: One potential issue is alcohol consumption, which is known to increase the risk of liver cancer. Smokers might also drink more, making it harder to determine if smoking alone causes liver cancer. So, I would say that alcohol is a confounding factor in this study.

Part B

A study found that teenagers who smoke are more likely to be depressed as young adults. Suggested causal relationship: Smoking causes depression.

Answer: It can be stated that genetics could pose an issue in this study. Some people might have genes that make them prone to both smoking and depression, which could complicate the study's findings. Another issue could be hormones, which is known to cause mood swings in teenagers during puberty. Hence, potential confounding factors could be genetics and hormones

Part C

A study found that people who consume lots of artificial sweeteners in diet soda are more likely to be overweight than people who do not drink diet soda. Suggested causal relationship: Artificial sweeteners cause obesity.

Answer: People who consume lots of artificial sweeteners may also have other dietary and lifestyle factors that contribute to obesity. What if the sample randomly picked individuals that perform little to no exercise? What if the sample randomly choose individuals that consume higher caloric food? Therefore, the confounding factors could be lifestyle and diet.

Part D

Observational studies show that women on the pill have a higher incidence of cervical cancer. Suggested causal relationship: The pill causes cervical cancer.

Answer: One major confounding factor could be unprotected sexual intercourse. Women may be on the pill (oral contraceptives), but if their partner does not wear a condom, they could be exposed to sexual transmitted diseases. One S.T.D. that is known to increase the chance of cervical cancer is human papillomavirus (HPV).

Part E

A study showed that infants living in homes that have two or more dogs or cats are less likely than other babies to develop allergies. Suggested causal relationship: Living with pets as an infant reduces allergy incidence.

Answer: In addition to genetics and nutrition, environmental factors, such as the cleanliness of the home environment and exposure to allergens other than pet dander, could also confound the relationship between pet ownership and allergy development. Given this knowledge, the casual relationship cannot be fully accuarate.

Part F

A study of 100,000 people (published in the Feb. 2003 issue of the journal Sleep) reported that people who reported sleeping eight or more hours per night had a higher mortality rate than those who slept seven or fewer hours. Suggested causal relationship: Sleeping more causes death

Answer:

Varies (of course), per person. If an individual has underlying health issues (like family history of health problems) they may require more sleep; which could cause a higher mortality rate. Another more straight forward confounding factor is the age. The elderly are commonly known to sleep less than younger individuals which could contribute to their morality rate. Additionally, being older is a strong indicator of having higher mortality rate (we all eventually die).

Question 2

Identify the population, variable, and parameter of interest in the following scientific questions

Part A

Estimate the average number of class credits taken by freshman at OSU this quarter

- **Population:** Freshman at OSU.
- **Variable:** Number of class credits taken by each freshman.
- **Parameter:** The average number of class credits taken by all freshman.

Part B

Test whether the median body temperature of cats is equal to the median body temperature of dogs.

- **Population:** Dogs and cats.
- **Variable:** Body temperature.
- **Parameter:** The difference in median body temperatures between cats and dogs. Specifically, checking for whether the median body temperature is equal between dogs and cats.

Part C

Test whether the variance of IQ scores for kindergarteners in the US is 10

- **Population:** U.S. Kindergarteners
- **Variable:** IQ scores.
- **Parameter:** Testing against the null hypothesis which suggests the variance of IQ scores is equal to 100. The alternative is that it is not 100.

Part D

Estimate the 20th percentile weight of rats on a particular reduced calorie diet

- **Population:** Rats on the specific reduced calorie diet.
- **Variable:** Weight of rats.
- **Parameter:** The 20th percentile weight of rats on this particular diet.

Question 3

Give one-sentence answers for the following:

Part A

A population parameter tells us about a whole group, while a statistic tells us about a smaller part of that group, like a sample.

Part B

A sampling distribution for a statistic is like a list of that statistic's values when we calculate it from different small groups taken from a big group, which helps us see how the statistic can vary from group to group.

Question 4

Perform simulations in R (like in part 4d of R Lab 1) to assess the quantities in part i. of each setting below. BE SURE TO INCLUDE YOUR R CODE AS AN APPENDIX.

How many simulated datasets did you generate to get a good estimate of the quantity? How did you decide that this was enough?

Answer: Simulated 5000 datasets for each part to estimate quantities. This sample size was chosen to ensure clear and stable normal distributions, as increasing it further did not significantly affect the distribution shape, as observed in the previous simulations (as in part 4d of R Lab 1).

Part A

Part I

```r
# Set the random seed
set.seed(10109)
# Parameters
n <- 20
rate <- 1
sim <- 5000
# Simulate the sample means
sampMeans <- replicate(sim, mean(rexp(n, rate)))
# Calculate P(Xbar > 1.3)
probSim <- mean(sampMeans > 1.3)
probSim
```

```
[1] 0.0978
```

Part A

Part II

```
# Calculate prob using the Gamma distribution
probGamma <- 1 - pgamma(1.3 * n, shape = n, scale = 1 / rate)
probGamma
```

```
[1] 0.09682085
```

Part A

Part III

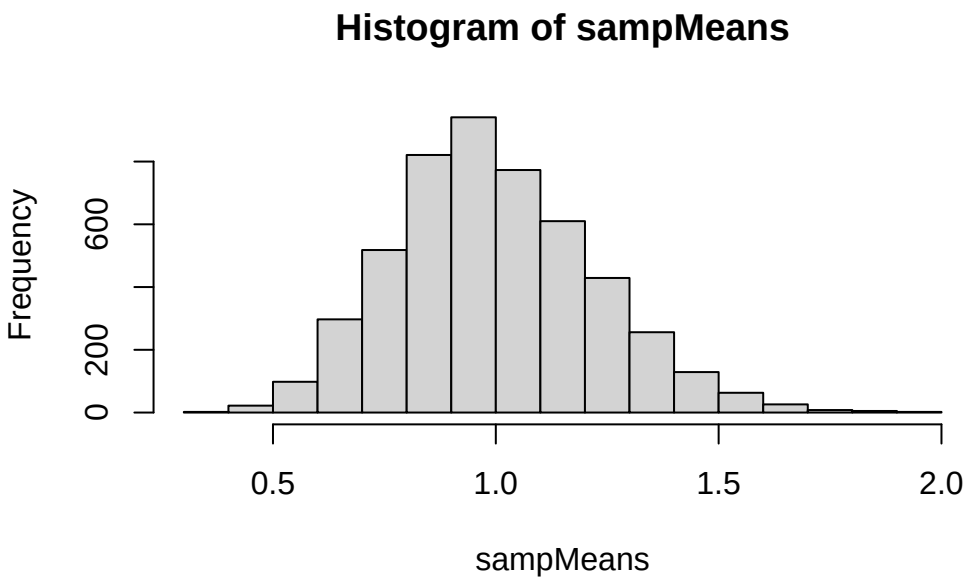
```
# Use the Central Limit Theorem to calculate the prob
mu <- rate
sigma <- rate / sqrt(n)
probCLT <- 1 - pnorm(1.3, mean = mu, sd = sigma)
probCLT
```

```
[1] 0.08985625
```

Part A

Part IV

```
hist(sampMeans)
```



The histogram is approximately bell-shaped, symmetrical distribution; hence, it indicates that the CLT is a reasonable approximation.

Part B

Part I

```
# Set the random seed
set.seed(10109)
# Set the parameters
n <- 5
max <- 1
min <- 0
sim <- 5000
# Simulate the sample means
sampMeans <- replicate(sim, mean(runif(n, min, max)))
# Calculate P(0.45 < Xbar < 0.55)
prob <- mean(sampMeans > 0.45 & sampMeans < 0.55)
prob
```

```
[1] 0.2902
```

Part B

Part II

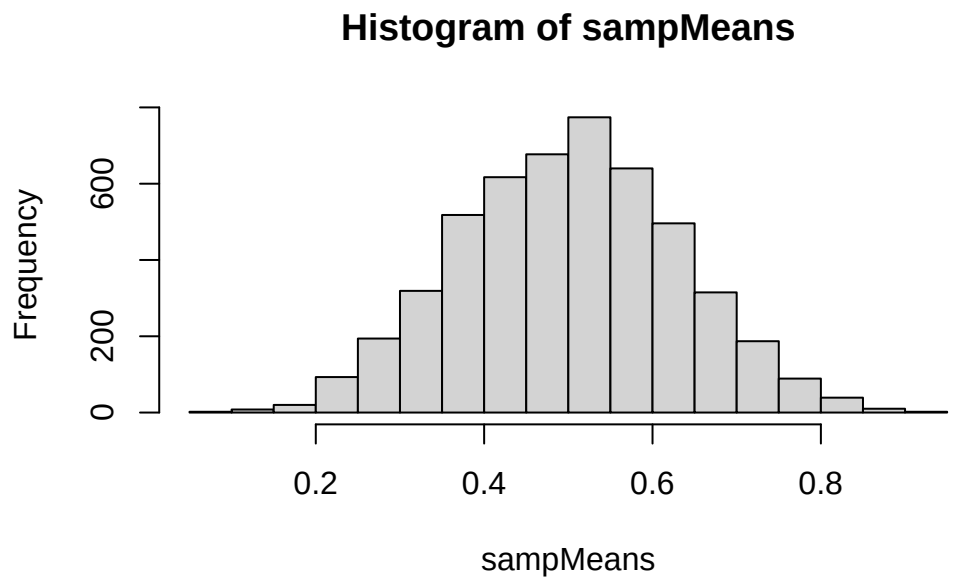
```
# Use the Central Limit Theorem to calculate the prob
mu <- (min + max) / 2
sigma <- (max - min) / (12 * sqrt(n))
probCLT <- pnorm(0.55, mean = mu, sd = sigma) - pnorm(0.45, mean = mu, sd = sigma)
probCLT
```

```
[1] 0.8202875
```

Part B

Part III

```
hist(sampMeans)
```



The histogram is, again, approximately bell-shaped, symmetrical distribution; hence, it indicates that the CLT is a reasonable approximation.

Part C

Part I

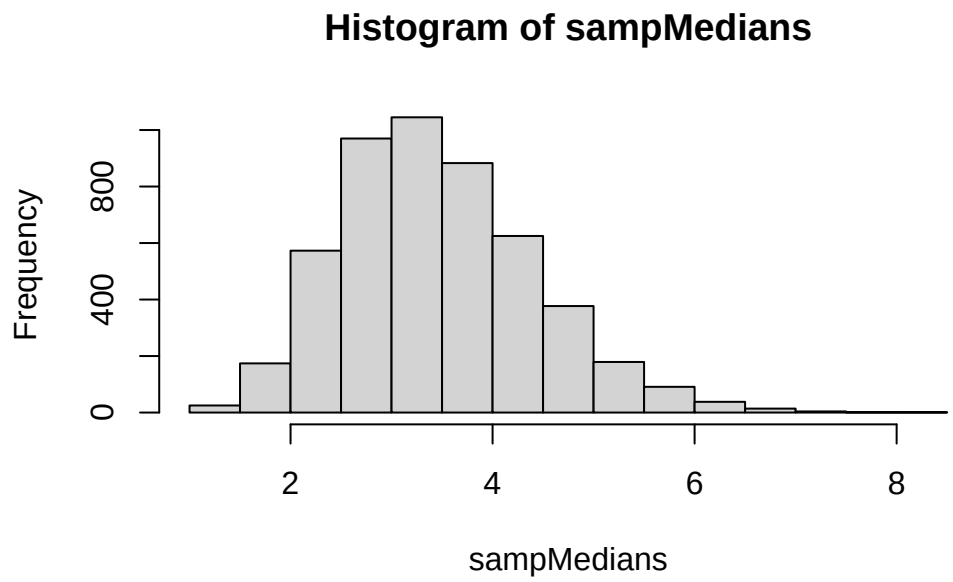
```
# Set the random seed
set.seed(10109)
# Set the parameters
n <- 10
degrees <- 4
sim <- 5000
# Simulate the sample medians
sampMedians <- replicate(sim, median(rchisq(n, degrees)))
# Calculate P(median > 4)
prob <- mean(sampMedians > 4)
prob
```

```
[1] 0.266
```

Part C

Part II

```
hist(sampMedians)
```



The histogram of sample means is right-skewed; however, despite the skewness in the individual data points, it still exhibits approximately a bell-shaped, symmetrical distribution. This suggests that the Central Limit Theorem is a reasonable approximation for the sample medians.

Part D

Part I

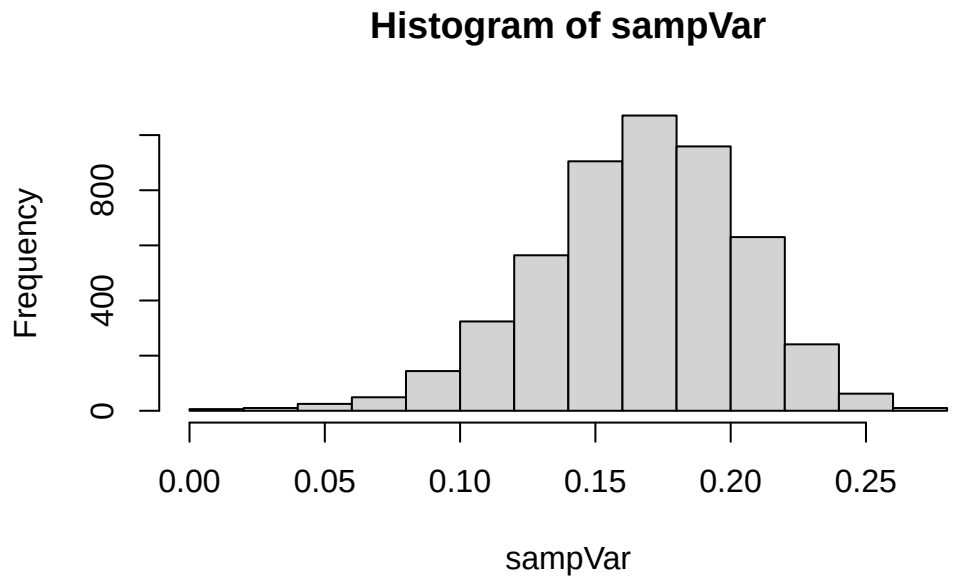
```
# Set the random seed
set.seed(10109)
# Set the parameters
n <- 10
shape1 <- 0.25
shape2 <- 0.25
sim <- 5000
# Simulate the sample variances
sampVar <- replicate(sim, var(rbeta(n, shape1, shape2)))
# Calculate P(variance > 0.2)
prob <- mean(sampVar > 0.2)
prob
```

```
[1] 0.1886
```

Part D

Part II

```
hist(sampVar)
```



The histogram of sample means is left-skewed; however, despite the skewness in the individual data points, it still exhibits approximately a bell-shaped, symmetrical distribution. This suggests that the Central Limit Theorem is a reasonable approximation for the sample variances.