ST 551 Homework 1

Student's Name Goes Here

Fall 2022

Instructions

This assignment is due by 11:59 PM, October 5th on Canvas via Gradescope. **You should submit your assignment as a typed PDF which you can compile using the provide .Rmd (R Markdown) template.** Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, grammatically correct sentences for your solutions.

Question 1 (18 points)

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 data set.
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.

(a) (3 points) Observational studies show that smokers have a higher rate of liver cancer. Suggested causal relationship: Smoking causes liver cancer.

(b) (3 points) A study found that teenagers who smoke are more likely to be depressed as young adults. Suggested causal relationship: Smoking causes depression.

(c) (3 points) 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.

(d) (3 points) Observational studies show that women on birth control pills have a higher incidences of cervical cancer. Suggested causal relationship: The pill causes cervical cancer.

(e) (3 points) 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.

(f) (3 points) 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.

Question 2 (12 points)

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

(a) (3 points) Estimate the average number of class credits taken by freshman at OSU this quarter.

- Population of interest:
- Variable of interest:
- Parameter of interest:

(b) (3 points) Test whether the median body temperature of cats is equal to the median body temperature of dogs.

- Population of interest:
- Variable of interest:
- Parameter of interest:

(c) (3 points) Test whether the variance of IQ scores for kindergartners in the US is 100.

- Population of interest:
- Variable of interest:
- Parameter of interest:

(d) (3 points) Estimate the 20th percentile weight of rats on a particular reduced calorie diet.

- Population of interest:
- Variable of interest:
- Parameter of interest:

Question 3 (4 points)

Give a one-sentence response for each of the following:

(a) (2 points) What is the difference between a population parameter and a statistic?

(b) (2 points) Describe what we mean by the *sampling distribution* for a statistic?

Question 4 (24 points)

Perform simulations in R to assess the quantities in part (i) of each population distribution below. Be sure to include the code used and try to output as few extraneous R outputs as possible (i.e. focus on outputting the answer but not the intermediate steps in your document).

a. (8 points) Population distribution: Exponential(rate = 1); Sample size: $n = 20$.

> **The R function `rexp(n, rate)` generates a sample of size `n` from the exponential distribution with rate parameter `rate`.**

(i) (2 points) Perform a simulation to assess $P(\bar{X} > 1.3)$ where $\bar{X}$ is the sample mean.

```r
# Write your code to answer this question below this line.
```

(ii) (2 points) Use the fact that the sum of n Exponenential(θ) random variables has a Gamma(n, θ) distribution to calculate the *exact* probability from part (i).

```r
# Use the pgamma(q, shape, rate) function to compute the exact probability.
# Hint: Here, shape = n and rate = theta
# Hint: pgamma(), by default, computes the probability of everything to the
#       left of the value q.
```

(iii) (2 points) Use the Central Limit Theorem to approximate the probability from part (i).

```r
# Use the pnorm(q, mean, sd) function to compute the approximate probability.
```

(iv) (2 points) Make a histogram of the simulated sample means you computed in part (i). Does is seem like the Central Limit Theorem would give a reasonable approximation?

```r
# Create a histogram of the means below this line
```

b. (6 points) Population distribution: Uniform(min = 0, max = 1); Sample size: $n = 5$**.**

The R function `runif(n, min, max)` generates a sample of size `n` from the uniform distribution ranging from value `min` to value `max`.

(i) (2 points) Perform a simulation to assess $P(0.45 < \bar{X} < 0.55)$ where $\bar{X}$ is the sample mean.

```r
# Write your code to answer this question below this line.
```

(ii) (2 points) Use the Central Limit Theorem to approximate the probability from part (i).

```r
# Use the pnorm(q, mean, sd) function to compute the approximate probability.
```

(iii) (2 points) Make a histogram of the simulated sample means. Does is seem like the Central Limit Theorem would give a reasonable approximation?

```r
# Create a histogram of the means below this line
```

c. (4 points) Population distribution: Chi-squared(df = 4); Sample size: $n = 10$.

The R function `rchisq(n, df)` generates a sample of size `n` from the chi-square distribution with degrees of freedom parameter `df`.

(i) (2 points) Perform a simulation to assess $P(m > 4)$ where m is the sample median.

```
# Write your code to answer this question below this line.
```

(ii) (2 points) Make a histogram of the simulated sample medians. Describe the *center*, *shape* and *spread* of the sampling distribution.

```
# Create a histogram of the medians below this line
```

d. (4 points) Population distribution: Beta(shape1 = 1/4, shape2 = 1/4); Sample size: $n = 10$**.**

The R function `rbeta(n, shape1, shape2)` generates a sample of size `n` from the beta distribution with shape parameter values `shape1` and `shape2`.

(i) (2 points) Perform a simulation to assess $P(s^2 > 0.2)$ where s^2 is the sample variance.

```
# Write your code to answer this question below this line
```

(ii) (2 points) Make a histogram of the simulated sample variances. Describe the *center*, *shape* and *spread* of the sampling distribution.

```
# Create a histogram of the variances below this line
```

e. How did you decide how many simulated data sets to generate in order to get a *good* estimate of the quantities above?