Lab Challenge 07 – Estimation

Due Date: 11:59 pm **the day before next class**
Each challenge is graded out of 2 points:

- 0 points – no attempt or no progress to a solution
- 1 point – challenge not fully completed or completed with major errors
- 2 points – challenge fully completed with at most a small error

Deliverables

1. A single pdf document containing your solutions to the challenges you completed.
2. An RStudio file (.R extension) containing a *complete* script used to generate your results.

Challenges

As in Lab 06, import the raw data for the variable named X.fail from the file "failures.txt". Each value of X.fail gives the number of hard drive errors detected in a large data center during one hour of operation over its entire history of operation (a period of several years).

1. Using the population data for X.fail, we can observe *why* the formula for sample variance uses $n-1$ instead of n.

$$s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2$$

   a. Import the raw population data for X.fail into R. Calculate its population variance σ^2. Note that R does not have a built-in function for population variance!

   b. In R, define two functions

```
s.var <- function (X) (1/(length(X)-1))*sum((X-mean(X))^2)
p.var <- function (X) (1/length(X))*sum((X-mean(X))^2)
```

   Simulate 10^5 samples of size $n = 5$ from the X.fail population data (without replacement). For each sample, evaluate both `s.var(this.sample)` and `p.var(this.sample)` and store those values as you go. Then calculate the mean of all the `s.var` values and all the `p.var` values.

   Which function, `s.var` or `p.var`, gives an unbiased estimate of the population variance?

2. Suppose a sample of 10 individual hours yields the sample data:

X.fail = 2, 2, 1, 0, 3, 1, 1, 1, 3, 2

For each of the following, round the limits of the confidence intervals to one decimal place.

a. Calculate a 95% confidence interval for $\mu =$ the population mean based on this sample.
b. Calculate a 90% confidence interval for $\mu =$ the population mean based on this sample.

3. Suppose each individual item in a large population of manufactured products is either defective (1) or non-defective (0). A random sample of $n = 300$ products were tested, yielding raw data:

Defect <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)

(These values are in the R script Defect.R.)

a. Using the above sample data, calculate a 99% confidence interval for the population proportion p of products that are defective.

b. Consider the following statement:

There is a 99% probability that p lies in the confidence interval from 3a.

This statement is wrong because it suggests that the true p varies randomly.

i. What is the thing that actually varies randomly in this situation?
ii. Rewrite the above sentence in a mathematically correct way.