Lab Challenge 06 – Sampling Distributions and Central Limit Theorem

Due Date: 11:59 pm, one week after 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

1. Assume that for each unit sold by your company, there is a 2.5% chance that the customer will require technical support within the first year of use. Now consider random samples of size $n = 400$ units. Let $\hat{p} =$ the proportion of units in the sample that will require technical support within the first year of use.
    a. Use R to plot a histogram showing the sampling distribution of $\hat{p}$. Use classes of width $\frac{1}{400} = 0.0025$. (You may use either simulation or find an exact result using dbinom.) Overlay a plot of the normal distribution with the same mean and standard deviation.
    b. Calculate the exact probability that $P(2.0\% \leq \hat{p} \leq 3.0\%)$.
    c. Calculate $P(2.0\% \leq \hat{p} \leq 3.0\%)$ using the normal approximation:
        i. without the "continuity correction"
        ii. with the "continuity correction"

For the next two challenges, you will need 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). This is the *population* data.

2. Using the raw data for the variable X.fail, complete the following.
    a. Plot a histogram of X.fail:
        - use `freq=FALSE` to show probability densities
        - use the `breaks` option to display one rectangle centered on 0, 1, 2, 3, etc., …
        - label the axes and give it a title
    b. Calculate the population mean and population standard deviation of X.fail.
    c. Use `lines()` to overlay a Poisson distribution on the histogram from part a.
        - Use the same mean (λ) as for X.fail.
        - Use options `type="p", pch="x", col="red"`

3. Use R to simulate selecting 10^5 samples of size $n = 50$ from this population (X.fail).
    a. Plot a histogram of the resulting sample means, $\bar{X}$. Use:
        - freq=FALSE
        - use the `breaks` option to get classes of width 0.1
    b. Calculate the mean and standard deviation of $\bar{X}$ based on the simulated samples.
    c. Use `lines()` to overlay a Normal distribution using:
        - the same μ and σ as the $\bar{X}$ distribution
        - use options `type="l", col="red"`
    d. What interval contains $\bar{X}$ for 95% of all samples of size $n = 50$? Round to two decimal places.