Homework 6

DUE: November 21 2022, 9PM

Instructions

Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `xtai_hw1.pdf`), to Gradescope (accessible through Canvas). You will give the commands to answer each question in its own code block, which will also produce output that will be automatically embedded in the output file. Each answer must be supported by written statements as well as any code used.

All code used to produce your results must be shown in your PDF file (e.g., do not use `echo = FALSE` or `include = FALSE` as options anywhere). `Rmd` files do not need to be submitted, but may be requested by the TA and must be available when the assignment is submitted.

Students may choose to collaborate with each other on the homework, but must clearly indicate with whom they collaborated.

Problem 1 (20 points)

Recall that in class (lecture 10), we learned the following rules of thumb for symmetric, bell-shaped distributions: 68%, 95%, and 99.7% of the values lie within one, two, and three standard deviations of the mean, respectively. This is actually derived from the normal distribution.

a. Use differences of two `pnorm()` values to derive each of the above three percentages.

b. For the two standard deviation version, do the same as in (a) using symmetry of the normal distribution (i.e., you should only use a single call of `pnorm()`).

Problem 2 (20 points)

Suppose $X_1, X_2, ..., X_n$ are independent $N(\mu, \sigma^2)$ random variables. Let $Y = \frac{\sum_{i=1}^{n} X_i}{2n}$.

a. What is the distribution of Y? Include information about the values of the population mean and variance.

b. Now assume n is large. Use the Central Limit Theorem to get an approximate distribution for Y, and derive its mean and variance.

Problem 3 (a-d: 30 points, e: bonus 10 points)

Assume that the number of branches on a redwood tree follows a normal distribution with mean 150 and standard deviation 30. Let the random variable X_i denote the number of branches on the ith redwood tree, where $i = 1, ..., n$. Then, $X_i \sim N(150, 30^2)$.

a. What is the probability of a tree having more than 180 branches? Calculate this in R using the original $X_i \sim N(150, 30^2)$ distribution, and after standardizing to a standard normal distribution.

b. Assume the samples are independent. What is the approximate distribution of the sampling distribution of the sample mean, $\overline{X}$?

c. Simulate 1000 draws, X_1 to X_{1000}, and calculate the sample mean.

d. Repeat (c) 5000 times and calculate the 5000 sample means. What are the mean and standard deviation of these 5000 sample means? Is it close to what you would expect?

e. (Bonus 10 points) If in (d) we repeated the experiment 10,000 times instead of 5,000 times, what would you expect the mean and standard deviation of the resulting sample means to be?

Problem 4 (30 points)

Assume that 70% of 18-20 year olds consume alcoholic beverages in any given year. Consider a random sample of 500 18-20 year olds. For each person, we record whether or not they have consumed alcoholic beverages in the past year.

a. Define a random variable X_i that describes the binary outcomes recorded (to be clear, we have 500 observations from this distribution). What distribution does your random variable follow? What is (are) the parameter(s)? What is the mean? What is the variance?

b. Calculating the fraction of 1's in our sample gives us a single observation from the sampling distribution of the sample proportion. What is the approximate distribution of this sampling distribution? Now, what is the approximate distribution of $\sum_{i=1}^{n} X_i$? Use this to calculate the probability that more than 361 students consumed alcoholic beverages.

c. A different way to think about $\sum_{i=1}^{n} X_i$ is a single draw from a different distribution. What distribution is this, and what are the parameters? Use R to calculate the exact probability that more than 361 students consumed alcoholic beverages. Is this different from your answer in (b)? Is this what you would expect? Please explain.

Appendix

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS  10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## loaded via a namespace (and not attached):
##  [1] compiler_4.0.2  magrittr_2.0.3  fastmap_1.1.0   cli_3.3.0
##  [5] tools_4.0.2     htmltools_0.5.2 rstudioapi_0.13 yaml_2.2.1
##  [9] stringi_1.7.8   rmarkdown_2.11  knitr_1.40      stringr_1.4.1
## [13] xfun_0.32       digest_0.6.25   rlang_1.0.4     evaluate_0.16
```