Homework 5

DUE: November 10 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 (25 points)

Consider a population in which 80% of those who are college-educated are employed, and 60% of those who are not college-educated are employed. In this population, 55% of individuals are not college-educated.

a. What is the probability of being employed? (You may handwrite your answer to this part, if you prefer. You can include an image in a code chunk using `knitr::include_graphics("myImg.png")`.)

b. If I pick five people at random from this population, what is the probability that none of those chosen is employed? (Hint: what random variable can we define? What distribution does this random variable follow?) Calculate the required probability by hand (you may use R as a calculator), then in R using a *single* function.

Problem 2 (45 points)

Assume that a college can admit at most 930 freshmen. Assume that it sends out 1500 acceptances and that each student comes to the college with probability .6, and that the students make decisions independently of one another.

a. What is the probability that the college ends up exactly the number of students it can accommodate?

b. What is the probability that the college ends up with with more students than it can accommodate?

c. What is the (theoretical) mean and variance of the distribution that you used in (a) and (b)?

d. In R, simulate the 1500 decisions that the accepted students make (i.e., create a binary vector of length 1500, indicating whether or not students attended the college). How many students, out of 1500, attended the college in your simulation? This number represents a single draw from the distribution that you used in (a) and (b).

e. In (d), what distribution did you use for each of the draws? What is (are) the parameter(s), and what is the theoretical mean and variance?

f. (Continued from (e)) As the sample size grows, what value do you expect the sample mean to converge to, and why? Does your answer in (d) make sense? If we had a sample size of 10000 (instead of 1500), what value would we expect for the number of students, out of 10000, that attend?

Problem 3 (25 points)

Assume that player ratings in chess tournaments follow a symmetric, bell-shaped distribution with average 1600 and standard deviation 350.

a. What common probability distribution do player ratings follow, and what are the parameters?

b. A player with a rating of 2650 enters the tournament. What is the probability of a rating higher than this player?

c. What is the probability of ratings between 1200 and 1800?

Problem 4 (Part a: 5 points, b-d: optional bonus 50 points)

Assume the number of accidents at a busy intersection is three a month on average, and follows a distribution commonly used to model rare events.

a. What distribution does the number of accidents at that intersection in a year follow, and what is (are) the parameter(s)? What is the mean and variance?

b-d (Bonus): We are going to illustrate the law of large numbers using the distribution that you have identified.

b. (Reminder: knowledge of this code is out of scope for this class, so this part is completely optional.) Below is code that we saw in class. First explain what the code does. Hint: look up the documentation for the `sapply()` function. The argument `...` represents optional arguments to `FUN`, in this case the second argument of the custom function `meanFun()`.

```
set.seed(0) # so results are reproducible (no need to explain this line)
binomDraws <- rbinom(n = 5000, size = 3, prob = .2)
myMeans <- data.frame(sampleSize = 1:5000, myMean = NA)
meanFun <- function(inputSampSize, outcomes) {
  return(mean(outcomes[1:inputSampSize]))
}
myMeans$myMean <- sapply(myMeans$sampleSize, meanFun, binomDraws)
head(myMeans)
```

c. Now copy the above code and change the relevant line(s) to sample from the distribution that you identified in (a).

d. What value do you expect the sample mean to converge to? Verify this using a line graph.

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
##
## other attached packages:
## [1] dplyr_1.0.9   ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
##  [1] pillar_1.8.1     compiler_4.0.2   tools_4.0.2      digest_0.6.25
##  [5] evaluate_0.16    lifecycle_1.0.1  tibble_3.1.8     gtable_0.3.0
##  [9] pkgconfig_2.0.3  rlang_1.0.4      cli_3.3.0        DBI_1.1.0
## [13] rstudioapi_0.13  yaml_2.2.1       xfun_0.32        fastmap_1.1.0
## [17] withr_2.4.2      stringr_1.4.1    knitr_1.40       generics_0.1.3
## [21] vctrs_0.4.1      grid_4.0.2       tidyselect_1.1.2 glue_1.6.2
## [25] R6_2.4.1         fansi_0.4.1      rmarkdown_2.11   blob_1.2.1
## [29] purrr_0.3.4      magrittr_2.0.3   scales_1.1.1     htmltools_0.5.2
## [33] assertthat_0.2.1 colorspace_1.4-1 utf8_1.1.4       stringi_1.7.8
## [37] munsell_0.5.0
```