Homework 2

Biostatistics I

Due Noon (12pm) October 4, 2017

For this assignment you are able to use either base `R` or `tidyverse` functions. However, be sure to document your code well, including any packages loaded. Remember, as always, to have complete figures and answer each question fully and with complete sentences.

Question 1 - Two dice

In this question we consider traditional 6-sided die with numbers 1,2,3,4,5, and 6. Die A is fair, meaning that all numbers have equal chance of being thrown. Die B is not fair: all odd numbers have the same chance as each other, and all even numbers have the same chance as each other, but even numbers are twice as likely to be thrown as off numbers.

- Part A: Write down the probability mass function for each of the two dice.
- Part B: Write down the cumulative distribution function for each of the two dice.

Question 2 - Throwing a 7

Many games involve throwing two dice, and getting two numbers that add up to `7` is a great result. Here we will simulate this scenario for two unfair dice (dice B).

- Part A: From your answer to Question 1, what is the probability of throwing a seven when throwing two dice B?
- Part B: Use a binomial distribution to simulate throwing these two dice (remember we only care about getting a seven or not). How many `7`s do you get if you throw the two dice 10 times? 100 times? 1,000 times? How close is this to the probability you stated in part A? Write a sentence comparing these counts to the probability you stated in part A.
- Part C: Simulate again throwing the two dice 10 times. How many `7`s did you get this time? Repeat this simulation 500 times or more, and each time record the number of `7`s you get. Plot a histogram of your results, and write a sentence describing it.
- Part D: Repeat part C when throwing the two dice 100 times, and 1,000 times. Plot your results. How are these plots similar and different from the plot in part C?
- Part E: Imagine I told you I threw the two dice 100 times and got a `7` 76 times. Would you think that was strange? How many of your simulations had 76 or more `7`s? Using the PMF or CDF, what is the probability of getting 76 or more `7`s when throwing the unfair die 100 times? Based on your answer, do you think I may have used a different die than the one you have?

Question 3 - Tidying Data

These data are from one of the largest trials in cancer screening, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) (https://prevention.cancer.gov/major-programs/prostate-lung-colorectal). This trial randomized 76,685 men and 78,216 women to either receive standard of care or additional screening in each of the cancers. Here we focus on the cohort of male participants who were randomized to be screened for prostate cancer. A csv with 25 of the 38,340 men in this group can be found on Canvas. These men were screened for prostate abnormalities using two tests: a digital rectal exam (DRE) and prostate specific antigen (PSA). DRE screening should have occurred on years 0,1,2, and 3, while PSA screening should have occurred on years 0,1,2 3, 4, and 5. However, since these are real data, many values are missing.

As you can see from the dataset, this table is *wide*, meaning that `psa_level0` is the PSA measurement at year 0 and `psa_level1` is the PSA measurement at year 1. It is your job to wrangle this dataset to a *long* format, with a column for `year`, one for `psa_level`, and one for `dre_result`. Be sure to order the final dataset by id and by year, and keep all clinical variables as well.

One approach is to separate the initial data into a `dat_demographics`, `dat_psa`, and `dat_dre`, making each dataset long individually, then finally combining them at the end. For the adventurous, it is definitely possible to avoid splitting the data using only the functions described in class.

Question 4 - Mean Squared Error

Show mathematically that the mean squared error of an estimator is equal to the sum of its variance and bias squared:

$$MSE[\hat{\theta}] = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

How does this property help us when defining the "best" estimator using MLE?

Question 5 - Maximum Likelihood Estimator

Consider the following i.i.d. sample from an Exponential(θ) distribution:

x
2.5303718
1.7298308
3.9871646
0.0947321
0.1686329
0.9495036
0.9426819
0.4358004
8.1787094
0.0874603
3.0144902
1.4406442
0.8430409
1.1313535
0.5648521
2.5493584
4.6896106
1.4362812
1.7728045
12.1230351

- Part A: Write the likelihood function of the exponential model
- Part B: Plot the likelihood function using the data provided above
- Part C: Compute the MLE of theta (provide both the equation and the actual numerical value with R). Be sure to show your work for both.