P8130 Fall 2020: Biostatistical Methods I

**Homework 2**

Due, Oct 7 @ 5:00pm

**P8130 Guidelines for Submitting Homework**

Your homework should be submitted only through CourseWorks. No email submissions!

All derivations, graphs, output and interpretations to each section of the problem(s) must be included in the PDF (not the code), otherwise it will not be graded.

Only 1 PDF file should be submitted. When derivations were required and handwriting was allowed, scan the derivations and merge ALL PDF files (http://www.pdfmerge.com/) into a single one.

We are encouraged to use R for calculations, but you still have to show the mathematical formulae. Also, make sure to also submit your commented code as a separate R/RMD file.

DO NOT FORGET:

You are encouraged to collectively look for answers, explain things to each other, and use questions to test each other knowledge.

*But*

Do NOT hand out answers to someone who has not done any work. Everyone ought to have ideas about the possible answers or at least some thoughts about how to probe the problem further. Write your own solutions!

Problem 1 (15p)

A new test is being developed for the detection of carcinoma of the prostate (*Foti et al.* N Engl. J Med. 1977). When it is tested in a group of 113 patients with prostatic cancer, 79 have a positive diagnosis; in a group of 217 individuals without prostatic cancer, 10 have a positive diagnosis.

1) Calculate the specificity and sensitivity of the test. (4p)

2) In another hypothetical scenario, it is planned to use the test to screen a large sample of subjects for prostatic cancer where the test results will be the only data available. Is this information enough to assess the test characteristics, i.e., sensitivity & specificity? Yes, no, what is missing (if the case)? (2p)

3) In patients with palpable prostatic nodules, the pretest likelihood of prostatic cancer is 50%. The test under these conditions has a sensitivity of 80% and a specificity of 95%.

a) Calculate the probability of a patient with a palpable prostatic nodule and a positive test result having prostatic cancer. What is the epidemiological term of this probability? (3.5p)

b) Re-calculate the probability in 2-a), using a pretest likelihood of prostatic cancer of 10%. Compare the two values and comment on their differences. (3.5p)

Problem 2 (10p)

According to the Center for Disease Control (CDC), about 34.5% of the adult US population are prediabetic (National Diabetes Statistics Report, 2020). Suppose we randomly select a group of 50 patients seen at Columbia University Medical Center.

Calculate the following exact probabilities based on the national statistics:

1) Probability that none of these patients are prediabetic. (2.5p)
2) Probability that less than 10 patients are prediabetic. (2.5p)
3) Probability that 34.5% of these patients are prediabetic (round to the nearest integer). (2.5p)
4) Could you use an approximation method to calculate the probabilities above? If yes, calculate the probabilities using approximations and compare to the exact values; otherwise, explain why approximations methods are not appropriate. (2.5p)

Problem 3 (10p)

The incidence of uveal melanoma in the US is approximately 5 per million individuals per year, with a significantly higher incidence in non-Hispanic Whites (6.02 per million), when compared to Blacks and Asians: 0.31 and 0.39 per million, respectively.

a) What is the probability that in NYC (population of 8.3 million reported in 2020), exactly 30 cases occur in a given year? (4p)
b) Compute the same probability in a) by the mentioned racial/ethnic groups and comment on the findings. Demographic data of NYC in 2020: 14.0% Asians, 42.8% non-Hispanic Whites, 24.3% Black. (6p)

Problem 4 (10p)

People with COVID-19 have had a wide range of symptoms, but high temperature is one of the most important indications. Based on current data, the average temperature of patients diagnosed with COVID-19 follow a normal distribution with a mean of 99.9 degrees Fahrenheit and a standard deviation of 0.73 degrees Fahrenheit. Of course, this differs from person to person, based on factors like body weight, height, the weather, age or gender.

Let $X_1, \dots, X_{40}$ be the body temperatures of 40 randomly chosen individuals returning to the Columbia University Medical Center. Calculate the following probabilities:

1) $P(\bar{X} < 98)$, the probability that the sample mean is less than 98 (average normal temperature). (2.5p)
2) $P(\bar{X} > 100.5)$, the probability that the sample mean is greater than 100.5 (alarming zone for COVID-19). (2.5p)
3) The 90$^{th}$ quantile of the sampling distribution of the sample mean $\bar{X}$. (2.5p)
4) The cutoff values for the middle 50% of the sampling distribution of the sample mean $\bar{X}$. (2.5p)