P8130: Biostatistical Methods I

Homework 1

Due, Sept 29th @ 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/SAS for calculations, but you still have to show the mathematical formulae. Also, make sure you include your commented code at the end of the document (PDF).

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

You are NOT supposed to 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 (10p)

A recent screening study for hypertension was performed among adults under age of 60 who are selected from the Washington Heights community. Out of 6,000 adults that were screened, 650 were found to have a diastolic blood pressure (DBP) of 90mmHg or higher. A random sample of 100 adults from those with DBP 90mmHg or higher and another random sample of 100 adults from those with DBP less than 90mmHg were sent to further investigations, and 75 and 15 of the respective samples were confirmed as hypertensive.

Extract the relevant information from the problem and answer the following questions. Where the case, also provide the clinical/epidemiological term of that specific probability.

a) What is the probability that an adult having DBP ≥ 90mmHg at screening will actually be hypertensive? (2p)
b) What is the probability that an adult having DBP < 90mmHg at screening will not actually be hypertensive? (2p)
c) What is the probability that an adult in this community is truly hypertensive? (2p)
d) What is the probability that a hypertensive person will be found to have a DBP ≥ 90mmHg at the initial screening? (2p)
e) What is the probability that a non-hypertensive person will be found to have a DBP < 90mmHg at the initial screening? (2p)

Problem 2 (5p)

It was found that that 8% percent of men and 30% of women cannot distinguish between colors red and green. This problem is called colorblindness and can cause problems with traffic signals.

a) If 10 men were randomly selected for a study on traffic signal perceptions, find the probability that exactly half of them cannot distinguish between red and green. (2.5p)
b) What if we would randomly select 10 women instead? How would the probability change? (2.5p)

Problem 3 (10p)

Records from state health departments showed that in early 1990s there were 9.35 cases of tuberculosis per 100,000 population in the US.

a) In a county of 60,000 people, what is the probability of observing 10 or more cases if the US rate held in this county? (2.5p)
b) What is the probability of 5 or fewer cases if the US rate held in this county? (2.5p)
c) Find the expected value and variance of this distribution function and interpret them in the context of the problem. (2.5p)
d) Could you calculate probabilities a) & b) using an approximation? (2.5p)

Problem 4 (10p)

The average time for a dental cleaning procedure is 35 minutes, with a variance of 25. Assume that the distribution of duration time for this procedure is approximately normal.

a) What value is 1 standard deviation above the mean? 1 standard deviation below the mean? What values are 2 standard deviations away from the mean? (2.5p)
b) Approximately what percentage of the times are between 25 and 45 minutes? (2.5p)
c) Approximately what percentage of the times are either less than 20 minutes or greater than 50 minutes? (2.5p)
d) Apply a z-transformation to this distribution and derive the expected value and variance for this new random variable. (2.5p)

Problem 5 (12p)

In a previous lecture, we discussed the concepts of theoretical vs empirical probabilities. Below is the illustrative example that was provided:

"In a class, 4 out of 10 students are allergic to peanuts. If you perform an experiment choosing 4 random students from this group, without replacement, what is the probability that they will all have peanuts allergy?"

We calculated the theoretical probability to be approximately 0.5%.

Use R/SAS to create your own simulations for calculating the empirical probabilities for the following scenarios:

a) Assume a constant 0.40 probability, simulate 10 trials for each of four students and calculate the percentage of times ALL four have peanuts allergy.
   Repeat this process for 100 trials and 5,000 trials and report your findings. (5p)
b) Generate one set of data for the entire group of 10 students. Note that the probability of allergies is 0.40 in this string of size 10. Select 10 random samples of size 4 without replacement, and record in how many instances ALL four students have peanuts allergy. Repeat the sampling process from the same data for 100 samples and 5,000 samples and report your findings. (5p)
c) Briefly comment on your findings and draw an overall conclusion. (2p)

Hint: R functions such as *rbinom()* and/or *sample()* might be useful. Also, make sure your findings are reproducible.