Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 5
Due: Thursday, October 13

1. Let Φ denote the CDF of the standard normal distribution. For each of the following, express your answer as a formula in terms of Φ and then evaluate numerically.
   (a) Area under the standard normal curve between -1.1 and -0.35
   (b) Probability that X lies outside the interval $[-2, 1]$ given that X follows a normal distribution with $\mu = -3$ and $\sigma = 3$
   (c) The area under the standard normal curve above _____ is 10%
   (d) The probability that X is below _____ is 25%, given that X follows a standard normal with $\mu = 100$ and $\sigma = 15$

2. This question is based on the NHANES sample of adult males available on the course website. For (a)-(c), **provide two answers**: the answer based on the normal approximation and the answer calculated directly from the data.
   (a) What percent of men weigh between 150 and 200 pounds?
   (b) What is the 25th percentile of weight?
   (c) How many men in the sample weigh above 300 pounds?
   (d) These approximations do not seem to be as good as those we saw in class. Why?

3. A study at Boston University found that for men who will develop coronary artery disease, cholesterol levels are normally distributed with a mean of 244 mg/dl and a standard deviation of 51 mg/dl. They also found that for men who do not develop the disease, cholesterol levels are normally distributed with a mean of 219 mg/dl and standard deviation 41 mg/dl. Consider the following "test" for coronary artery disease: if a man has cholesterol level above 240, we predict that he will develop coronary artery disease (*i.e.*, he tests positive).
   (a) What is the sensitivity of this screening tool?
   (b) What is the specificity of this screening tool?
   (c) If ten individuals who will not develop coronary artery disease take this test, what is the probability of obtaining at least 2 false positives?

4. A researcher wants her sample mean to be twice as accurate; how much does she have to increase her sample size by?

5. An article in the *New England Journal of Medicine* reported that among adults living in the United States, the average level of albumin in cerebrospinal fluid is 29.5 mg/dl, with a standard deviation of 9.25 mg/dl. We are going to select a sample of size 20 from this population.
   (a) How does the variability of our sample mean compare with the variability of albumin levels in the population?

(b) What is the probability that our sample mean will be greater than 33 mg/dl?

(c) What is the probability that our sample mean will lie between 29 and 31 mg/dl?

(d) What two values will contain the middle 50% of our sample means?

6. According to an article in the *American Journal of Public Health*, the distribution of birth weights in a certain population is approximately normal with mean 3500 grams and standard deviation 430 grams.

   (a) What is the probability that a newborn's weight will be less than 3200 grams?

   (b) Suppose we take a sample of 9 newborns. What is the probability that their average weight will be less than 3200 grams?

   (c) In the aforementioned sample of 9 newborns, how many newborns would you expect to weigh under 3200 grams?

   (d) What is the probability that our sample of 9 newborns will contain exactly 3 newborns who weigh less than 3200 grams?

   (e) Suppose we take 5 samples of 9 newborns. What is the probability that at least one of the sample averages will be less than 3200 grams?

   (f) How large must our sample be in order to ensure a 95% probability that the sample mean will be within 50 grams of the population mean?

7. On the course website is a data set called `nhanes-subsamples.txt`. Each column of the data set contains 1,000 sample means of triglyceride levels calculated from 1,000 randomly drawn subsamples. For each column, however, the number of women in those subsamples was different. For each of the three columns (A, B, and C), how large was the sample size? All of the sample sizes I used are multiples of ten, so please select your answers from $\{10, 20, 30, \ldots\}$. Also, please describe how you came to your answers.

8. A common test for association between a trait and a gene in genetics is the transmission disequilibrium test. It relies on finding (i.e., sampling) parent-child pairs in which the child has the trait of interest and the parent is heterozygous for the gene of interest (i.e., has one copy of each version of the gene). The parent is equally likely to pass on either copy, so if there is no link between the trait and the gene, we would expect 50% of the children to have version "A" and the other 50% to have version "B". However, because we have systematically sampled only children with the trait (any children without the trait are not included in the study), if version "A" causes a child to be more likely to develop the trait of interest, we would expect to find a higher proportion of version "A" in the children than version "B". In other words, if the two are associated, the "transmission" of the gene is distorted away from "equilibrium", hence the name of the test.

   In a 1989 study reported in the journal *Genetic epidemiology*, data was collected for 124 such parent-child pairs, in which the offspring had Type I diabetes and the parent was heterozygous for 5'FP (a flanking polymorphism adjacent to the insulin gene on chromosome 11). Among the children, 78 received the "class 1" version of 5'FP from their parent, while the other 46 did not.

   (a) Carry out the transmission disequilibrium test based on the exact distribution of the number of offspring who received the "class 1" version of 5'FP. What is the probability of seeing a distortion as extreme or more extreme than the 78/46 split we saw in the data, if there really is no link between 5'FP and Type I diabetes?

(b) Carry out the transmission disequilibrium test based on the central limit theorem approximation for the number of offspring who received the "class 1" version of 5'FP. As in part (a), what is the probability of seeing a distortion as extreme or more extreme than the one observed if 5'FP and Type I diabetes truly are independent?

(c) In the *Genetic epidemiology* article, the authors report the results from the approximate test in (b) and state that "an exact binomial test can be used, if desired, instead". Expand on their advice. Does the difference between the approximate and exact test matter in this study? If not, when might it matter?

9. A crossover study published in the *American Journal of Clinical Nutrition* investigated whether oat bran cereal helps to lower serum cholesterol levels. Fourteen individuals with high cholesterol were placed on a diet that included either oat bran or corn flakes; after two weeks, their LDL cholesterol levels were measured. Each individual was then switched to the other diet; after two weeks, the LDL levels were recorded again. The data from the study are on the course website. Let π denote the probability that an individual from this population would would experience a reduction in cholesterol on an oat bran diet vs. a corn flake diet.

   (a) Construct the 95% Clopper-Pearson interval for π

   (b) Construct the 95% Wald interval for π

   (c) Construct the 95% score interval π

10. In assignment 4, we analyzed data from a paired study of treatments for Parkinson's disease in which 50 out of 78 pairs did better on deep-brain stimulation than the control. Let π denote the probability that an individual with Parkinson's disease would benefit from deep-brain stimulation.

    (a) Construct the 95% Clopper-Pearson interval for π

    (b) Construct the 95% Wald interval for π

    (c) Construct the 95% score interval π

    (d) These intervals agree with each other much better than the three intervals in the previous problem; why?

11. In lab we conducted a simulation looking at the coverage of the Clopper-Pearson interval as a function of π. Instead, carry out a simulation examining the coverage of the Clopper-Pearson interval as a function of n, keeping π fixed at 0.25. Choose an "interesting" sequence of values for n (e.g., don't choose 9, 10, and 11). Produce a plot similar to the one we produced in lab, where coverage is plotted vs. n, and place a horizontal line at 0.95 to represent the nominal coverage of the interval.