Biostatistical Methods I (BIOS 5710)
Breheny

Assignment 6
Due: Thursday, October 20

1. In lab, we ran a simulation in which we looked at the sampling distribution of the mean for samples of size 25 from the NHANES study containing the triglyceride levels of 3,026 adult women.

   (a) Repeat the simulation for samples of size 5 instead, and provide a histogram of the sampling distribution.

   (b) Repeat the simulation for samples of size 100 instead, and provide a histogram of the sampling distribution.

   (c) Comment briefly on the differences between the histograms in (a), (b), and the histogram we obtained in lab.

2. True or false:

   (a) To create a confidence interval for the mean of a sample with 5 observations, we would use the t distribution with 5 degrees of freedom.

   (b) The area outside $\pm$ 2 for the t distribution with 10 degrees of freedom will be larger than the area outside $\pm$ 2 for the t distribution with 50 degrees of freedom.

   (c) The z-test and t-test will be in closer agreement when n is small than when n is large.

   (d) When performing a t-test, I need to worry more about the distribution of the data when n is small than when n is large.

3. (a) A paper in *Pediatrics* reported on a sample of 10 infants receiving a certain type of antacid to treat digestive disorders. The antacids contained aluminum, and physicians were concerned about the levels of aluminum in the plasma of these infants. In the sample, the mean aluminum level was 37.2 μg/l, with standard deviation 7.13 μg/l. Calculate a 95% confidence interval for the average plasma aluminum level of infants taking these antacids.

   (b) The average plasma level of aluminum in the general population of infants is between 4 and 5 μg/l. What would you conclude about the levels of aluminum in infants taking these antacids versus the average levels in healthy infants?

4. In population A, the standard deviation of LDL cholesterol is 20 mg/dL, and in population B, the standard deviation is 40 mg/dL. An investigator collects random samples of 10 individuals and uses the t distribution to calculate a 95% confidence interval from each population. Which of the following is true: (i) the confidence interval for population A will definitely be wider (ii) the confidence interval for population A will probably be wider (iii) the two confidence intervals should be about equally wide (iv) the confidence interval for population B will probably be wider (v) the confidence interval for population B will definitely be wider

5. People with diabetes often sustain vascular damage to their retina (diabetic retinopathy), which can lead to blindness. Levels of vascular endothelial growth factor (VEGF) in the eye are highly correlated with retinopathy, and often used as a marker of the severity of the damage. In one study

published in the *New England Journal of Medicine*, researchers performed a treatment called laser photocoagulation on seven patients with diabetic retinopathy. They measured the ocular VEGF levels before and after treatment and found that VEGF levels were lower by an average of 4.2 ng/mL, with a standard deviation of 2.85 ng/mL.

(a) What is wrong with the following argument? "4.2 is only about one and a half standard deviations away from 0. Differences need to be 2 or more standard deviations away in order to be significant. Therefore, this evidence is not statistically significant."

(b) Construct a 95% confidence interval for the average decrease in ocular VEGF levels following treatment.

(c) Conduct a paired t-test of the hypothesis that the treatment has no effect on VEGF levels.

(d) Does laser photocoagulation seem to help or hurt patients with diabetic retinopathy? Or is difficult to say, because this study provides little evidence of an effect in either direction?

6. This problem involves the same crossover study of oat bran diets and serum cholesterol levels that was in assignment 5.

(a) Perform a paired t-test of the hypothesis that oat bran consumption has no effect on serum cholesterol.

(b) Construct a 95% confidence interval for the average amount by which an oat bran diet could be expected to reduce a man's serum LDL cholesterol levels (as compared to a corn flake diet).

(c) Compared with the binomial test, name an advantage of the paired t-test.

(d) Compared with the binomial test, name a disadvantage of the paired t-test.

7. (a) Suppose that X follows a χ^2_ν distribution. Show that $\mathrm{E}(X) = \nu$.

(b) Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent and follow a $\mathrm{N}(\mu, \sigma^2)$ distribution. Show that S^2 is an unbiased estimator of σ^2.[1]

8. The *mean squared error* of an estimator $\hat{\theta}$ is defined as

$$\mathrm{MSE} = \mathrm{E}\{(\hat{\theta} - \theta)^2\};$$

i.e., we're looking at the error of an estimator $\hat{\theta} - \theta$, squaring that, then taking the mean/expected value. Show that

$$\mathrm{MSE} = \mathrm{Bias}^2 + \mathrm{Var}(\hat{\theta}),$$

where $\mathrm{Bias} = \mathrm{E}(\hat{\theta}) - \theta$. Hint: Try adding and subtracting $\mathrm{E}(\hat{\theta})$ from the term being squared, then expanding the square.

9. In lab, we saw that the confidence interval derived from the pivotal quantity $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ did not hold up well for non-normally distributed data with $n = 25$. Does the accuracy of this interval improve (i.e., get closer to the nominal level) as n gets larger? Try this out with a simulation and then explain why the accuracy of the interval is improving, or why it isn't. For this simulation study, set `replace=TRUE` when using `sample`; this doesn't make much difference when n is small, but if you increase n into the thousands, it makes a big difference. If `replace=FALSE` and n is large, the sample is artificially close to the population (because our population here is fairly small) and coverage is artificially inflated.

[1] This is true even if X does not follow a normal distribution, but for the purposes of this problem you only need to show it for normally distributed random variables.