

**BST 140.651**

**Problem Set 3**

**Due in class on October 13th**

Problem 1. Much discussion has appeared in the medical literature in recent years on the role of diet in the development of heart disease. The serum cholesterol levels of a group of persons who eat a primarily macrobiotic diet are measured and it is found that among 24 persons ages 20-39, the sample mean cholesterol is 220 with a sample standard deviation of 35.

- a. If the mean cholesterol level in the general population in this age group is 230 and is assumed to be normally distributed, then test the hypothesis that the group of persons on a macrobiotic diet have lower cholesterol levels than the general population. Find the p-value. Interpret.
- b. Compute a 95% confidence interval for the true mean cholesterol for people who eat the macrobiotic diet.

Problem 2. Referring to problem 1 a. above, suppose a larger study is now being planned. We anticipate recruiting 100 subjects on the macrobiotic diet.

- a. Compute the power of a one-sided test with significance level .05 if the true mean serum cholesterol for this group is 225. Do the same calculation allowing the true mean serum cholesterol for this group to be 223 and 220.
- b. Sketch the power curve (that is, plot power versus the alternative).

Problem 3. The level of serum creatinine in the blood is measured in 15 drug users. The creatinine levels are

0.9	0.7	1.0
1.1	1.4	1.1
1.6	1.2	1.4
2.0	1.5	2.2
0.8	0.8	1.4

Suppose the mean serum creatinine in the general population is 1.0. Perform a hypothesis test on the above data to determine if there is evidence that creatinine levels for drug users are different than the general population. Compute a 90% confidence interval for the mean level among drug users. (You must state the hypotheses, present p-values, interpret your results and state your assumptions.)

Problem 4. A random sample was taken of 20 patients admitted to a hospital with a certain diagnosis. The lengths of stays in days for the 20 patients were

4,	2,	4,	7,	1,	5,	3,	2,	2,	4
5,	2,	5,	3,	1,	4,	3,	1,	1,	3

- a. Calculate a 95% confidence interval (use the method  $\bar{X} \pm t \text{ SE}$ ) for the mean length of hospital stay. Is your answer reasonable? What underlying assumptions were required for this method and are they reasonable?
- b. Calculate a 95% confidence interval using bootstrapping.
- c. Take logs of the data (base "e")
  - (i) Calculate a 95% confidence interval for the mean of the log length of stay.
  - (ii) Take antilogs (exponentiate) of the endpoints of the confidence interval found in part (C(i)). Explain why that is a 95% confidence interval for the median length of stay if the data is lognormally distributed (lognormally distributed is when the logarithm of the data points has a normal distribution). Technically, under the lognormal assumption, is the confidence interval you found in this equation also a confidence interval for the mean length of stay?

Problem 5. Let  $p$  denote the unknown proportion of rocks in a riverbed that are sedimentary in type. Suppose that  $X = 12$  of a sample of  $n = 20$  rocks collected in random locations are found to be sedimentary in type.

- a. Plot the likelihood for the parameter  $p$ .
- b. From your graphs, determine the value of  $\hat{p}$  of  $p$  where the curve reaches its maximum. Does this value for the maximum make intuitive sense? Comment in one or two sentences.
- c. Show that the point that maximizes the binomial likelihood is always  $X/n$ .

Problem 6. The 'spleen rate' in a population is of great interest to malaria epidemiologists. MacDonald (The Epidemiology and Control of Malaria, 1957) defines this as: 'The percentage of children aged from 2 to 10 inclusive, in whom the spleen is palpable when in the standing position.'

Suppose you palpate the abdomen of 85 randomly selected children, aged 2-10, in a malarious area. In 12 of the children the spleen is palpable.

- a. Calculate the maximum likelihood estimate for the true 'spleen rate'. List the assumptions that you use in this calculation.
- b. Calculate 95% and 99% confidence intervals for the true 'spleen rate' in the population from which these children were drawn.
- c. Plot the likelihood for the true spleen rate. Draw 1/8 and 1/32 reference lines.

Problem 7. A group of investigators wish to explore the relationship between the use of hair dyes and the development of breast cancer in females. A group of 1000 beauticians 40-49 years of age is identified and is followed for 5 years at which time it is discovered that 20 new cases of breast cancer have occurred. Let us assume breast cancer incidence over this time period for an average American woman in this age group is 7/1000. We wish to test the hypothesis that use of hair dyes increases the risk of breast cancer.

- a. State the hypothesis, perform an analysis, report a  $p$ -value and interpret.
- b. Plot the associated likelihood. Draw  $1/8$  and  $1/32$  reference lines.
- c. Comment in a few (3 or 4) sentences on the relative merits of the design of this study. Specifically is it prospective or retrospective? Is there a control group?

Problem 8. This item investigates the performance of the Wald confidence interval

- a. Using a computer, generate 1000 Binomial random variables for  $n = 10$  and  $p = .3$
- b. Calculate the percentage of times that

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1 - \hat{p})/n}$$

contains the true value of  $p$ . Here  $\hat{p} = X/n$  where  $X$  is each binomial variable.

- c. Do the intervals appear to have the coverage that they are supposed to?
- d. Repeat the calculation only now use the interval

$$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1 - \tilde{p})/n}$$

where  $\tilde{p} = (X + 2)/(n + 4)$ . Does the coverage appear to be closer to .95?

- e. Repeat this comparison (parts a. - d.) for  $p = .1$  and  $p = .5$ . Which of the two intervals appears to perform better? Why?
- f. Extra credit: [5 points] Plot the coverage for the two definitions of the binomial confidence interval for the true value of  $p$  ranging between 0 and 1. (Consider at least 100 values of  $p$  between 0 and 1.) Draw a horizontal reference line at .95. Which of the two intervals appears to perform better across the range of values for  $p$ ? You may not ask the instructor or TAs about the extra credit problem.