BST 140.651-652
Problem Set 4

Problem 1. A special study is conducted to test the hypothesis that persons with glaucoma have higher blood pressure than average. Two hundred subjects with glaucoma are recruited with a sample mean systolic blood pressure of $140mm$ and a sample standard deviation of $25mm$. (Do not use a computer for this problem.)

a. Construct a 95% confidence interval for the mean systolic blood pressure among persons with glaucoma. Do you need to assume normality? Explain.

b. If the average systolic blood pressure for persons without glaucoma of comparable age is $130mm$. Is there statistical evidence that the blood pressure is elevated?

Problem 2. Consider the previous question.

a. Make a probabilistic argument that the interval

$$\left[\bar X - z_{.95}\frac{s}{\sqrt{n}}, ~~\infty\right]$$

is a 95% *lower bound* for μ.

Problem 3. Suppose we wish to estimate the concentration μg/mℓ of a specific dose of ampicillin in the urine. We recruit 25 volunteers and find that they have sample mean concentration of 7.0 μg/m ℓ with sample standard deviation 3.0 μ g/mℓ. Let us assume that the underlying population distribution of concentrations is normally distributed.

a. Find a 90% confidence interval for the population mean concentration.

b. How large a sample would be needed to insure that the length of the confidence interval is 0.5 μ g/mℓ if it is assumed that the sample standard deviation remains at 3.0 μ g/m ℓ?

Problem 4. Here we will verify that standardized means of iid normal data follow Gossett's t distribution. Randomly generate $1,000 \times 20$ normals with mean 5 and variance 2. Place these results in a matrix with $1,000$ rows. Using two `apply` statements on the matrix, create two vectors, one of the sample mean from each row and one of the sample standard deviation from each row. From these $1,000$ means and standard deviations, create $1,000$ t statistics. Now use R's `rt` function to directly generate $1,000$ t random variables with 19 df. Use R's `qqplot` function to plot the quantiles of the constructed t random variables versus R's t random variables. Do the quantiles agree? Describe why they should.

Problem 5. Here we will verify the chi-squared result. Simulate $1,000$ sample variances of 20 observations from a normal distribution with mean 5 and variance 2. Convert these sample variances so that they should be chi-squared random variables with 19 degrees of freedom. Now simulate $1,000$ random chi-squared variables with 19 degrees of freedom using R's `rchisq` function. Use R's `qqplot` function to plot the quantiles of the constructed chi-squared random variables versus those of R's random chi-squared variables. Do the quantiles agree? Describe why they should.

Problem 6. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then we know that $(n-1)S^2/\sigma^2$ is chi-squared with $n-1$ degrees of freedom. You were told that the expected value of a chi-squared is its degrees of freedom. Use this fact to verify the (already known fact) that $E[S^2] = \sigma^2$. (Note that S^2 is unbiased holds regardless of whether or not the data are normally distributed. Here we are just showing that the chi-squared result for normal data is not a contradiction of unbiasednes.)

Problem 7. A study of blood alcohol levels (mg/100 ml) at post mortem examination from traffic accident victims involved taking one blood sample from the leg, A, and another from the heart, B. The results were:

Case	A	B	Case	A	B
1	44	44	11	265	277
2	265	269	12	27	39
3	250	256	13	68	84
4	153	154	14	230	228
5	88	83	15	180	187
6	180	185	16	149	155
7	35	36	17	286	290
8	494	502	18	72	80
9	249	249	19	39	50
10	204	208	20	272	271

a. Create a graphical display comparing a case's blood alcohol level in the heart to that in the leg. Comment on any interesting patterns from your graph.

b. Create a graphical display of the distribution of the difference in blood alcohol levels between the heart and the leg.

c. Do these results indicate that in general blood alcohol levels may differ between samples taken from the leg and the heart? Give confidence interval and interpret your results.

d. Create a profile likelihood for the true mean difference and interpret.

e. Create a likelihood for the variance of the difference in alcohol levels and interpret.

Problem 8. 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 $\overline{X} \pm t \quad$ 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% percentile bootstrap interval and interpret.

Problem 9. Refer to the previous problem. Take logs of the data (base "e")

a. Calculate a 95% confidence interval for the mean of the log length of stay.

b. Take antilogs (exponential) of the endpoints of the confidence interval found in part a.. 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 that you found in this equation also a confidence interval for the mean length of stay?

Problem 10. 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 and interpret.

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.

d. Use the CLT to create a confidence interval for the true proportion of rocks that are sedimentary. Interpret your results.

e. A much larger study is planned and the researchers would like to know how large n should be to have a margin of error on the estimate for the proportion of sedimentary rocks that is no larger than $.01$ for a 95% confidence interval? Use the fact that $p(1-p) \leq 1/4$. Also try the calculation with the estimate of p from the current study.

Problem 11. This problem investigates the performance of the Wald confidence interval

a. Using a computer, generate 1000 Binomial random variables for $n = 10$ and $p = .3$ 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. Do the intervals appear to have the coverage that they are supposed to?

b. 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?

c. Repeat this comparison (parts a. - d.) for $p = .1$ and $p = .5$. Which of the two intervals appears to perform better?

Problem 12. Forced expiratory volume FEV is a standard measure of pulmonary function. We would expect that any reasonable measure of pulmonary function would reflect the fact that a person's pulmonary function declines with age after age 20. Suppose we test this hypothesis by looking at 10 nonsmoking males ages 35-39, heights 68-72 inches and measure their FEV initially and then once again 2 years later. We obtain this data.

Person	Year 0 FEV (L)	Year 2 FEV (L)	Person	Year 0 FEV (L)	Year 2 FEV (L)
1	3.22	2.95	6	3.25	3.20
2	4.06	3.75	7	4.20	3.90
3	3.85	4.00	8	3.05	2.76
4	3.50	3.42	9	2.86	2.75
5	2.80	2.77	10	3.50	3.32

a. Create the relevant confidence interval and interpret.

b. Create the relevant profile likelihood and interpret.

c. Create a likelihood function for the variance of the change in FEV.

Problem 13. Another aspect of the preceding study involves looking at the effect of smoking on baseline pulmonary function and on change in pulmonary function over time. We must be careful since FEV depends on many factors, particularly age and height. Suppose we have a comparable group of 15 men in the same age and height group who are smokers and we measure their FEV at year 0. The data are given (For purposes of this exercise assume equal variance where appropriate).

Person	FEV Year 0 (L)	FEV Year 2 (L)	Person	FEV Year 0 (L)	FEV Year 2 (L)
1	2.85	2.88	9	2.76	3.02
2	3.32	3.40	10	3.00	3.08
3	3.01	3.02	11	3.26	3.00
4	2.95	2.84	12	2.84	3.40
5	2.78	2.75	13	2.50	2.59
6	2.86	3.20	14	3.59	3.29
7	2.78	2.96	15	3.30	3.32
8	2.90	2.74			

a. Graphically display the distribution of change in pulmonary function for smokers and nonsmokers (from the previous problem).

b. Calculate a confidence interval to determine if there is evidence to suggest that the change in pulmonary function over 2 years is the same in the two groups. State your assumptions and interpret your results.

Problem 14. In a trial to compare a stannous fluoride dentifrice A, with a commericially available fluoride free dentifrice D, 260 children received A and 289 received D for a 3-year period. The mean DMFS increments (the number of new Decayed Missing and Filled tooth Surfaces) were 9.78 with standard deviation 7.51 for A and 12.83 with standard deviation 8.31 for D. Is this good evidence that, in general, one of these dentifrices is better than the other at reducing tooth decay? If so, within what limits would the average annual difference in DMFS increment be expected to be?

Problem 15. Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects' body mass indices (BMIs) were measured at a baseline and again after having received the treatment or placebo for four weeks. The average difference from follow-up to the baseline (followup - baseline) was $-3\ kg/m^2$ for the treated group and $1\ kg/m^2$ for the placebo group. The corresponding standard deviations of the differences was $1.5\ kg/m^2$ for the treatment group and $1.8\ kg/m^2$ for the placebo group. Does the change in BMI over the two year period appear to differ between the treated and placebo groups? (Show some work and interpret your results.) Assume normality and a common variance.

Problem 16. In a random sample of 100 subjects with low back pain, 27 reported an improvement in symptoms after execise therapy. Give and interpret an interval estimate for the true proportion of subjects who respond to exercise therapy.

Problem 17. Suppose that systolic blood pressures were taken on 16 oral contraceptive users and 16 controls at baseline and again then two years later. The average difference from follow-up SBP to the baseline (followup - baseline) was $11\ mmHg$ for oral contraceptive users and $4\ mmHg$ for controls. The corresponding standard deviations of the differences was $20\ mmHg$ for OC users and $28\ mmHg$ for controls.

a. Calculate and interpret a 95% confidence interval for the change in systolic blood pressure for oral contraceptive users; assume normality.

b. Does the change in SBP over the two year period appear to differ between oral contraceptive users and controls? Create the relevant 95% confidence interval and interpret. Assume normality and a common variance.