STA 2101/442 Assignment Four[1]

Please bring your R printouts for questions 2, 3 and 11e to the quiz. Please print them separately, because you may be asked to hand in just one of them. The other questions are just practice for the quiz, and are not to be handed in, though you may use R as a calculator. Bring a real calculator to the quiz.

1. The statistic
$$W_n = n\,(\mathbf{L}\mathbf{T}_n - \mathbf{h})' \left(\mathbf{L}\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}'\right)^{-1} (\mathbf{L}\mathbf{T}_n - \mathbf{h})$$
provides a large-sample test of $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$. Let $X_1, \ldots, X_n$ be a random sample from a $B(1, \theta)$ distribution.

   (a) Write down and simplify the W_n statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

   (b) Your answer is related to one of the test statistics Z_1 and Z_2 introduced earlier for testing this same null hypothesis. To which one is W_n related, and how is it related?

2. In Assignment One, you tested difference between means for the Verbal SAT and the Math SAT score. A link to the SAT data is available from the course home page. Using R, please calculate the W_n statistic to test this hypothesis. Feel free to use my code directly. Note that the statistic $\mathbf{T}_n$ is of dimension *two*. Guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Be able to state your conclusion in plain, non-statistical language. Bring your R printout to the quiz.

3. Suppose you want to test for differences between *variances* of Verbal SAT and Math SAT, again with minimal assumptions about their joint distribution. The idea is to first convince yourself that the joint distribution of $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ is asymptotically bivariate normal. Then, if you can get a good estimate of their asymptotic covariance matrix, you can use the W_n test. For simplicity, let $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ have n in the denominator rather than $n - 1$.

   (a) Write down a formula for one of the sample variances, say $\widehat{\sigma}_1^2$, in a form that shows how it's a continuous function of a collection of sample means.

   (b) Letting X_i denote performance on the Verbal SAT and Y_i denote performance on the Math SAT, we want a data vector $\mathbf{D}_i$ to which we can apply the Central Limit Theorem. Then, we would write $\widehat{\sigma}_1^2 - \widehat{\sigma}_2^2 = g(\overline{\mathbf{D}}_n)$ and the delta method would establish asymptotic normality. Show you know what's going on by writing down the data vector $\mathbf{D}_i$.

   (c) It's too much work to calculate the Jacobian and then estimate all the moments in the asymptotic covariance matrix, so we'll use the Bootstrap instead. Let's all use a bootstrap sample size of $B = 1,000$. Using R, calculate an estimated asymptotic covariance matrix.

[1]Copyright information is at the end of the last page.

(d) Using R, please calculate the W_n statistic to test $H_0 : \sigma_1^2 = \sigma_2^2$.

(e) Guided by the usual $\alpha = 0.05$ significance level, what do you conclude? Be able to state your conclusion in plain, non-statistical language.

Bring your R printout to the quiz.

4. A team of botanists grew fungus in a nutrient solution in test tubes. Each day for seven days, one of their graduate students carefully measured the length of the fungus in each of n tubes. The scientists were interested in lots of things, including whether average growth was linear or not. Denote the expected amount of fungus at day j by μ_j.

   (a) What is the null hypothesis, in symbols?

   (b) Assuming that the scientists wish to make as few assumptions as possible and n is large, the W_n statistic is natural for this problem. What is $\mathbf{T}_n$?

   (c) What is $\mathbf{L}$?

   (d) What is $\mathbf{h}$?

   (e) What is a convenient choice for $\widehat{\mathbf{\Sigma}}_n$? How many rows and columns?

5. Ten friends have a party right after graduating from university. At the time, none of them has ever been married. The party includes a visit by a fortune teller, who says "Five years from now, 3 of you will still be unmarried, 3 of you will be married for the first time, 2 will be divorced, one will be married for the second time, and one will be widowed." How many ways are there for this to happen? The answer is a number. Show your work.

6. A fair die is tossed 8 times. What is the probability of observing the numbers 3 and 4 twice each, and the others once each? The answer is a number.

7. A box contains 5 red, 3 white and two blue marbles. A sample of six marbles is drawn with replacement. Find the probability that

   (a) 3 are red, 2 are white and one is blue

   (b) 2 are red, 3 are white and 1 is blue

   (c) 2 of each colour appears.

   All the answers are numbers.

8. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a $M\left(1, (\theta_1, \ldots, \theta_c)\right)$ distribution. Show why the likelihood function is written $L(\boldsymbol{\theta}) = \theta_1^{n_1}\theta_2^{n_2}\cdots\theta_c^{n_c}$.

9. Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a random sample from a $M\left(1, (\theta_1, \theta_2, \theta_3)\right)$ distribution. Find the maximum likelihood estimator of $(\theta_1, \theta_2, \theta_3)$. Show *all* your work.

10. Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution with expected value μ and variance one, and independently, let $Y_1, \ldots, Y_n$ be another random sample from the same distribution (with the same n). Find the asymptotic distribution of $\frac{\overline{X}_n}{\overline{Y}_n}$. What happens when $\mu = 0$?

11. In a political poll, a random sample of registered voters indicate which party they generally like most: Conservative, NDP or Liberal (other preferences were indicated by a small number of respondents; they are excluded from this analysis). A multinomial model seems reasonable for these data, with n_1, n_2 and n_3 denoting the number who chose Conservative, NDP and Liberal respectively. Of course $n = n_1 + n_2 + n_3$.

    The *odds* of an event is the probability of the event divided by one minus the probability. Take the natural log and you have the *log odds*, a quantity that has a prominent role in categorical data analysis.

    (a) Give consistent estimators of the log odds of supporting the Conservatives and the log odds of supporting the NDP. How do you know the estimators are consistent?

    (b) Find the approximate large sample *joint* distribution of the two log odds estimators. Show your work. The covariance matrix has a fairly nice form.

    (c) Express your answer to the last part by saying "They're approximately bivariate normal (what else?) with expected value ..."

    (d) Suppose that in a random sample of 200 voters, 91 chose the Conservatives, 71 the NDP and 38 the liberals. Give

        i. The estimated asymptotic covariance matrix of the estimators. Your answer is a 3×3 matrix of numbers. Show your work.

        ii. A point estimate of odds (not log odds) of choosing the NDP. The answer is one number.

        iii. a 95% confidence interval for the log odds of choosing the NDP. The answer is a pair of numbers.

        iv. Using your answer to the last part (the accepted way to do it), give a 95% confidence interval for the odds (not log odds) of choosing the NDP. The answer is a pair of numbers.

    (e) In Question 11d, your estimated asymptotic covariance matrix was based on the delta method. Produce another estimated asymptotic covariance matrix using the bootstrap. Let's all use a bootstrap sample size of $B = 1,000$. Bring your R printout to the quiz. I used the `rmultinom` function to do my bootstrap. See `help(rmultinom)` and think about it.

    Here's some output from two independent runs of my bootstrap. They are close to my delta method estimate. Please do *not* use my crazy names `logoddscon` and `logoddsndp`.

```
            logoddscon  logoddsndp
logoddscon  0.02051701 -0.01420638
logoddsndp -0.01420638  0.02196846

            logoddscon  logoddsndp
logoddscon  0.02032482 -0.01431120
logoddsndp -0.01431120  0.02160154
```