STA 2101/442 Assignment Five[1]

Please bring your R printouts to the quiz. The non-computer 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. Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution. Data are given in the file `shoes.data`. There is a link on the course web page in case the one in this document does not work.

   As you may or may not know, the sampling distribution of the sample median is asymptotically normal, with asymptotic mean equal to the population median. Using this fact, obtain an approximate 95% confidence interval for the median of the unknown distribution. Bring your R printout to the quiz and be ready to explain what you did.

2. For a random sample from a Multinomial$(1, \boldsymbol{\theta})$ distribution, show that the likelihood ratio test statistic can be written as

$$G^2 = 2n \sum_{j=1}^{c} \overline{Y}_j \log\left(\frac{\overline{Y}_j}{\widehat{\theta}_j}\right),$$

   where $\widehat{\theta}_j$ is the *restricted* maximum likelihood estimate of θ_j. That is, it's the MLE under H_0. You may use without proof the fact that the unrestricted MLE is $\overline{Y}_j$.

3. In yet another marketing study, consumers are given samples of three different laundry detergents, which we will call "A," "B" and "C." The labels that the people see are randomly assigned. Suppose $n = 150$ consumers participate in the product test. Thirty-eight prefer "A", 55 prefer "B,", and 57 prefer "C."

   (a) Test the null hypothesis of equal preference using a large-sample likelihood ratio test. Of course you may use Question 2.

   (b) Test the same null hypothesis using a Wald test.

   In both cases, use `R`. Your printout should display the value of the test statistic, the degrees of freedom and the p-value. Do you reject H_0 at the 0.05 significance level? Bring your printout to the quiz.

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

4. Based on a random sample of size n from a p-dimensional multivariate normal distribution, derive a formula for the large-sample likelihood ratio test statistic G^2 for the null hypothesis that $\mathbf{\Sigma}$ is diagonal (all covariances between variables are zero). You may use material on the slides from the multivariate normal lecture (including a useful version of the likelihood function), which will be provided with the quiz if this question is asked. You may also use without proof the fact that the unrestricted MLE is $\widehat{\theta} = (\overline{\mathbf{x}}, \widehat{\mathbf{\Sigma}})$, where

$$\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'.$$

Hint: Because zero covariance implies independence for the multivariate normal, the joint density is a product of marginals under H_0.

5. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are given in the file `sat.data`. There is a link on the course web page in case the one in this document does not work.

   Using R, carry out a large-sample likelihood ratio test to determine whether there are any non-zero covariances among the three variables; use $\alpha = 0.05$.

   (a) Calculate the test statistic G^2 and the p-value; also give the degrees of freedom. Your answer is a set of 3 numbers.

   (b) Do you reject H_0? Answer Yes or No.

   (c) Are the three variables all independent of one another? Answer Yes or No.

   Bring your printout to the quiz.

6. Men and women are calling a technical support line according to independent Poisson processes with rates λ_1 and λ_2 per hour. Data for 144 hours are available, but unfortunately the sex of the caller was not recorded. All we have is the number of callers for each hour, which is distributed Poisson$(\lambda_1 + \lambda_2)$. Here are the data, which are also available in the file `poisson.data` on the course website:

```
12  9 25 11 22 16 17  8 14 17 14 17  8 14 18 16 13 17 13 11  9 15 18
14 16 17 16 19 13 18 12 12 13 10  8 15 13 11 15 15  6 10 13 13 11 13
 9 15 16  9  5 10  8 18 13 17  7 13 13 17 12 17 14 16  6 12 17 10  9
14 11 19 13 17 15 20 14 10 13 14 17  9 13 14  7 16 16  9 25 10 10  9
17  7 15 12 14 21 14 18 14 12 13 15 12 11 16 14 15 16  8 19 13 17 15
11 18 13 12 11 19 14 16 17 13 13 19 19 11 19 10 12  9 18 11 14  9 14
14 14 13  9 13 18
```

(a) The parameter in this problem is $\boldsymbol{\theta} = (\lambda_1, \lambda_2)'$. Try to find the MLE analytically. Show your work. Are there any points in the parameter space where both partial derivatives are zero?

(b) Now try to find the MLE numerically with R's `nlm` function. The Hessian is interesting; ask for it. Try two different starting values. Compare the minus log likelihoods at your two answers. What seems to be happening here?

(c) Try inverting the Hessian to get the asymptotic covariance matrix. Any comments?

(d) To understand what happened in the last item, calculate the Fisher information in a single observation from the definition. That is, letting $\ell = \log f(Y; \boldsymbol{\theta})$, calculate the elements of the 2×2 matrix whose (i, j) element is

$$-E\left(\frac{\partial^2 \ell}{\partial\theta_i \partial\theta_j}\right)$$

(e) The Fisher information in the sample is just n times the Fisher information in a single observation. Using the numerical MLEs from one of your `nlm` runs, estimate this quantity (a 2×2 matrix). Compare it to the Hessian. Now do you see what happened when you tried to calculate the asymptotic covariance matrix?

Bring your printout to the quiz.

Most good homework problems have a lesson. The lesson here is that it's possible for a model to be perfectly be correct and the sample size to be large, but the data are still not adequate to allow successful estimation of the model parameters by maximum likelihood (or, it turns out, by any other method[2]). And the main clue is a Hessian matrix that is not positive definite.

---

[2]When two or more sets of parameter values give rise to exactly the same probability distribution for the observed data, using the data to decide which one is correct is a hopeless task, and all reasonable methods of estimation will fail. In such cases, the parameter vector is said to be *not identifiable*.