University of Western Ontario
Statistics 3858b Final Exam

April 16, 2012

Instructor: R. J. Kulperger

Instructions:

I. Make sure that your name and ID number are the front of your exam booklet. This exam is 3 hours in length, from 9 AM to 12 PM (12 Noon).

II. Additional handouts for the exam : a formula sheet, and the normal, chi-square distribution, student's t distribution and Mann-Whitney tables from the text.

NAME : _____

ID : _____

1	2	3	4	5	total

1. (a) Define *statistic*. Give a statistical model and an example of a r.v. that is **not a statistic**.

   (b) State the Neyman-Pearson Lemma. Discuss in one or two sentences why this is important in statistical theory.

   (c) State the Cramer-Rao Lower bound. Discuss in one or two sentences why this is important in statistical theory.

   (d) State the Rao-Blackwell Theorem. Discuss in one or two sentences why this is important in statistical theory.

2. Consider the following Bayes estimation problem.

   Suppose $X_i, i = 1, \ldots, n$ are iid, conditional on $\Theta = \theta$, exponential, mean $\frac{1}{\theta}$

   Let f_Θ be the prior distribution and suppose it is Gamma(α, β).

   (a) Find the posterior distribution of Θ, that is $f_{\Theta|X_1,\ldots,X_n}(.|x_1, \ldots, x_n)$ the conditional pdf of $\Theta|X_1 = x_1, \ldots, X_n = x_n$. *Hint: You might find it easiest to just work with the kernel, that is something that is proportional the posterior pdf, and then at the end find the normalizing constant*

   (b) Find the Bayes estimator of Θ. *Aside : It will not matter here if you use Bayes estimator or Bayes estimate.*

   (c) Is the prior distribution used here a conjugate prior? You will give Yes or No with a one or two sentence explanation.

3. In class we studied two nonparametric methods.

(a) Consider the nonparametric bootstrap

(i) for r.v.s X_i , $i = 1, \ldots, n$ give empirical distribution function, call this F_n

(ii) For data 1.8 , -1, 0.5, 3 give formula for F_n and graph it.

(iii) Give a method to simulate a r.v. X^* that has cdf F_n. *Hint : It will involve either a continuous or discrete uniform distribution.*

You may answer this generally or for the specific F_n in the previous part, whichever you wish. If you give the method for the F_n in the previous part, then indicate how this generalizes to other sample sizes n.

(iv) The r.v.

$$W = \frac{\sqrt{n}(\bar{X} - \mathrm{E}(X))}{\sqrt{S^2}}$$

can be used to obtain a confidence interval for $\mu = \mathrm{E}(X)$, and where S^2 is the sample variance. If X_i are iid normal we can use the student's t distribution to obtain quantiles. The student t quantiles are often not correct if the population distribution is not normal.

How can one use the nonparametric bootstrap to obtain a confidence interval for μ. Give the algorithm (no proofs needed anywhere in this) to obtain the confidence interval for μ.

The page is also available for your work. Part b of this question, dealing with the rank statistic, Mann-Whitney, is on the page following that.

(b) Consider two independent random samples, $X_i, i = 1, \ldots, n$ are iid with cdf F and $Y_j, j = 1, \ldots, m$ are iid with cdf G. Suppose also that F, G are continuous.

Let Y_Y be the rank sum statistic for the random variables $Y_1, \ldots, Y_m$.

i. Suppose $F = G$. Obtain, with appropriate steps in your calculations to justify your answer, $E(T_Y)$.

ii. Now consider the setting with $n = 4, m = 2$.
The $\binom{6}{2}$ subsets of size 2 are given in Table 1, given on the next page. There is also a column for your convenience to record the rank sum for the subsets of size 2.
Assume $F = G$. Obtain $P(T_Y \geq 10)$ and $P(|T_Y - 7| \geq 3)$.

iii. Consider $H_0 : F = G$ versus the alternative $H_A : F \neq G$.
Observe data $x_1 = 14, x_2 = 18.1, x_3 = 6, x_4 = 12, y_1 = 2, y_2 = 8$.
Give the observed value of T_Y, the rank sum of y_1, y_2.

Table 1: Table 1: Subsets of Size 2

subsets		rank sum
1	2	
1	3	
1	4	
1	5	
1	6	
2	3	
2	4	
2	5	
2	6	
3	4	
3	5	
3	6	
4	5	
4	6	
5	6	

Give the p-value for this observed data. State your conclusion.

4. (a) Consider an iid sample $X_i, i = 1, \ldots, n$ from a distribution with pdf or pmf of the form $f(.;\theta)$ and where $\theta \in \Theta$, an appropriate parameter space. To be specific you may just work with the case of f being a pdf. State what is meant by a regular case for the pdf f, specifically giving the two conditions.

   (b) For a regular case prove, showing enough steps as needed, and with $X \sim f(.;\theta)$ that

$$\mathrm{E}_\theta \left(\frac{\partial \log(f(X;\theta))}{\partial \theta} \right) = 0$$

   and Fisher's information

$$I(\theta) = -\mathrm{E}_\theta \left(\frac{\partial^2 \log(f(X;\theta))}{\partial \theta^2} \right) .$$

   (c) In this and the remaining part of this question you will work with $f(.;\theta)$ being the exponential pdf, with mean $\frac{1}{\theta}$. $X_i, i = 1, \ldots, n$ is an iid sample from this model.

      i. Find the MLE of θ. Call this MLE $\hat{\theta}_n$. Also Find Fisher's information.

      ii. State an appropriate Theorem from our class that will give a limit distribution for the r..v

$$\sqrt{n}(\hat{\theta}_n - \theta) .$$

      Give this limit distribution for our MLE problem.

      iii. Suppose we have a sample of size $n = 64$ and some summary statistics of this data.

      Some summary statistics are

$$\sum_{i=1}^{64} x_i = 22.2714 , \sum_{i=1}^{64} x_i^2 = 12.45304 .$$

      Give the observed Fisher's information.

      Give the approximate 90% confidence interval for θ.

   (d) Using the MLE in the previous part determine if $\hat{\theta}_n$ is an unbiased or biased estimator.

   In either case determine if it is asymptotically unbiased.

   Also if the estimator is biased, can you find a simple *fix* to make it unbiased, and if so do this.

   (e) Find the method of moments estimator for θ. Is it the same r.v. as the MLE in this case?

5. An experiment is performed in which the data falls into one of 3 categories, say 1 , 2 and 3. The model is that each trial falls into j with probability $p_j = \binom{2}{j}(1-\theta)^{2-j}\theta^j$, where $\theta \in \Theta = [0,1]$. The experimenter performs n iid trials, with random outcomes X_i, $i = 1, \ldots, n$. He then records the

$$N = (N_0, N_1, N_2) = \left(\sum_{i=1}^{n} I(X_i = 0), \sum_{i=1}^{n} I(X_i = 1), \sum_{i=1}^{n} I(X_i = 2) \right)$$

which counts the number of observations that fall into the categories.

The experimenter will test the null hypothesis $H_0 : \theta = \frac{1}{2}$ versus the alternative $H_A : \theta > \frac{1}{2}$.

(a) Derive the Generalized Likelihood Ratio test Λ for this problem.

Specifically you will need to give the likelihood function, and the formula for the argmax in the numerator and denominator.

In this part you do not need to work with the rejection region, just derive the GLR.

(b) Use an appropriate Taylor's formula approximation to show that $-2\log(\Lambda)$ is approximated by Pearson's chi-square statistic. Specifically you will derive Pearson's chi-square statistic for this particular problem.

(c) State a Theorem that gives the limit distribution for $-2\log(\Lambda)$ (and hence for Pearson's chi-square), and give the degrees of freedom for this particular problem.

For size $\alpha = .01$ what is the rejection region, including the critical value (or values), for this test of hypothesis.

(d) The experimenter performs $n = 100$ independent experimental trials. The data is

$$N_{\text{observed}} = (23, 48, 29)$$

Give the value of the argmax for the denominator.

Give the observed value of the Pearson's chi-square statistic.

Perform the test of hypothesis at level or size $\alpha = .01$ and state your conclusion.