Math 30 – Mathematical Statistics

Final Exam - Takehome

Distributed May 6, 2011

Due May 13, 2011 by 5 p.m. ***sharp*** to 306 SM (or faxed/emailed)

Instructions:

1. Show all work (on your own paper or blue book). You may receive partial credit for partially completed problems, so you should try all parts of all problems. If you want a blue book, feel free to take one. Turn in ONLY your FINAL solutions please. Please write solutions legibly!
2. The exam is open book and open notes. You may refer to old homework and exams, as well as any handouts or other items posted on our course website.
3. You may not use any other references or other texts and no online media (i.e. closed general internet).
4. You may use a calculator, Excel, R, or any other statistical computing program to help with calculations. All calculations can be completed just with a calculator.
5. You may not discuss the exam with anyone but me until after the exam has been turned in by all students. If you ask me a clarification question, I will send my reply to all students in the class.
6. You may take as much time as you like and may complete the exam in multiple sittings.
7. Suggestion: Point values per problem are displayed below if that helps you allocate your time among problems.
8. Office hours: You will be able to find me in my office on Monday 9-12 (during other course exam), Tuesday 10-12, Wednesday 1-3, and Thursday 1-3. Also, on Friday just before the exam is due. For other times, just send me an email, and we'll arrange a time to meet for your questions.
9. Good luck!

Problem	1	2	3	4	5	6	7	Total
Possible Points	15	11	20	8	12	14	20	100

1. Hypothesis Testing and a Bit of Estimation (15)

Assume you take 2 independent random samples, each of size n, from two (possibly different) Poisson distributions. In other words, you have $X_1, X_2, \ldots, X_n$ from a Poisson (λ_X), and $Y_1, Y_2, \ldots, Y_n$ from a Poisson (λ_Y). Suppose you want to develop a test procedure to test

$H_0 : \lambda_X + \lambda_Y = s$ versus $H_A : \lambda_X + \lambda_Y \neq s$, where s>0 is a given positive constant.

a. Determine if the null and alternative are simple or composite hypotheses, and explain why.

b. Determine a test procedure to test the given hypotheses, clearly stating your test statistic, and giving a distribution for it (it may be asymptotic, not exact).

c. In this setting, you observe 5 observations for X, and 5 for Y (see table). Is there evidence that ***s*** is different from 15 at a .05 significance level?

X	10	7	11	9	6
Y	2	4	2	1	4

d. Suppose you thought you were in this situation and just wanted to estimate ***s***. What is the MLE estimate for ***s*** (formula and specific value for this data)?

2. Nonparametric Test Statistic Exact Distributions (11)

Suppose you observe X_1, X_2, X_3, and Y_1, Y_2, Y_3 as random observations from two independent populations and you want to test if ***X*** and ***Y*** have the same distribution or not in the 2-sample shift model (i.e. the only way the distributions are allowed to differ is in their location).

a. Find the ***exact*** null distributions of the Wilcoxon rank sum statistic, W, and the Mann-Whitney U statistic in this situation.

b. Use your distributions to test the relevant hypotheses for this data: X observed = 1.8, 6, 4.92, and Y observed = 17.93, 3.89, 18.40. Use a significance level no greater than .25 for this main test. Report your attained significance level as well.

c. Compare your exact results to what the normal approximation for the Mann-Whitney U test would give for your observed U and one other more extreme U value. In this situation where the conditions for the normal approximation are not met, does the normal model appear to be doing a good job or not?

3. Quadratic Regression with Constraints (20)

Suppose you sample n pairs of observations in a setting where you believe a regression model is appropriate and you believe that the relationship between Y and X is given by $E(Y_i) = \beta_1 x_i + \beta_2 x_i^2$, i=1,…,n. Additionally you know the setup of the experiment is such that $\sum x_i = \sum x_i^3 = 0$ (symmetric around 0). Finally, you assume the other regression assumptions are met – homoscedasticity, independence and normality of the error terms.

a. Determine the normal equations for the model and use them to solve for explicit forms for $\hat{\beta}_1$ and $\hat{\beta}_2$.

b. Fit the model to the data set displayed in the table below and report the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$.

c. Determine an unbiased estimator for σ^2 and give its formula for this setting and obtain its value for this data set. You do not need to prove that it is unbiased.

d. Obtain the residuals from your regression model. Use the Kolmogorov-Smirnov test to determine whether the assumption related to the distribution of the error terms (estimated by the residuals) is met (you will need to use your estimate from c.).

e. In class we found the expected values and variances of the parameter estimates for a regression setting with an intercept and a slope. Now, in this setting with 2 slopes and no intercept, and treating the x values as given or conditioned on (i.e. you may treat them as constants), prove that $\hat{\beta}_1$ and $\hat{\beta}_2$ are unbiased estimators. Then, find their variances as functions of σ^2. For this data set, which estimator, $\hat{\beta}_1$ or $\hat{\beta}_2$, will have the lowest variance?

Regression Data:

X	-4	-3	-2	-1	0	0	1	2	3	4
Y	107.62	52.98	23.56	1.30	1.15	2.80	20.05	38.82	89.56	151.88

4. Alien Candy (8)

On the planet Xenoxis, a new candy has just been introduced that has four colors, each with an associated flavor. A statistics class on Xenoxis wants to know if the distribution of colors follows the manufacturer's specified distribution, which they obtained (see table), but which includes an annoying parameter θ that changes daily. A random sample of 200 candies produced on one day yields the observed counts:

Color	1	2	3	4
Proportion	θ^2	θ	$\theta - 2\theta^2$	$(1-\theta)^2$
Observed Count	36	88	10	66

Perform a hypothesis test with a .05 significance level to see if the data is consistent with the null hypothesis that there exists a θ such that the proportions of each color category are as specified in the table. Provide an estimate of that θ (that makes sense given your data).

5. General Testing (12)

a. Perform 2 different tests (one parametric and one nonparametric) to determine if the population with sample observations below has a center (mean/median, etc.) at 40 or if the center is higher than 40. You may assume that the population is normally distributed. You should use a .025 significance level for each test (or as close as you can get without going over). Additionally, if you must pre-test before doing any test, use a .025 significance level for any pre-tests. A random sample was drawn and the sample observations were:

36.67	40.51	41.55	41.87	42.94	53.04
33.81	39.55	40.80	41.91	44.34	50.06

b. Find the power of the parametric test in this situation if the true mean is really 45. Would you expect a higher power if you increased n? Would you expect a higher power if the true mean was really 50? Explain.

6. Bayesian Inference from 2 Observations (14)

Suppose you observe X and Y, both of which follow a normal distribution with mean 0 and precision θ. (i.e. given θ, X and Y are N(0, σ) where $\sigma^2 = 1/\theta$). Suppose you are a Bayesian, so you assign a prior to θ as $g(\theta) = \lambda^2 \theta e^{-\lambda\theta}$, with $\theta > 0, \lambda > 0$.

a. With $\lambda = 1$, what (common) distribution does this prior resemble? For general λ, what (common) distribution does this prior resemble? In general, is it a proper prior? In general, is it an informative prior?

b. Determine the posterior distribution of θ, after observing X and Y.

c. Is this prior a conjugate prior? Explain.

d. Determine the Bayes' estimator for θ (formula).

e. Suppose you observe X =1.55 and Y=-.12. Use the data to determine a 95% posterior credible interval for θ if $\lambda = 2$. If you cannot obtain numerical endpoints for your interval, explain how you would form one, using notation for the missing critical values. (You will not be penalized for doing the critical value approach, as getting the endpoints requires appropriate transformations or using R/some other software).

7. Estimation Recap (20)

A random variable X is said to have a Pareto distribution with parameters x_0 and α (both > 0) if it has pdf given by $f(x \mid x_0, \alpha) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}, x \geq x_0$, and 0, otherwise. Suppose you have a random sample of observations $X_1, X_2, \ldots, X_n$ from a Pareto distribution.

a. Assume x_0 is known (or could be conditioned on or estimated), and find the MLE for α.
(Hint: consider a transformation of X to get rid of issues with α in the exponent).

b. Now suppose that α is known and x_0 is not. Determine the MLE of x_0.

c. Determine sufficient statistics for x_0 and α, in the settings of parts b and a respectively, and justify how you know they are sufficient. (I.E. identify a sufficient statistic for α when x_0 is known, and a sufficient statistic for x_0 when α is known.)

d. Are your estimators in parts a and b sufficient or minimal sufficient or neither? Explain.

e. You may take as fact that if α >1, $E(X) = \frac{\alpha x_0}{\alpha - 1}$, and that if α >2, $Var(X) = \frac{\alpha x_0^2}{(\alpha-1)^2(\alpha-2)}$.

i. If α is known and is >2, what is the MoM estimator of x_0 ? How does this compare to the MLE?

ii. Is the MoM estimator of x_0 unbiased? If so, show it. If not unbiased, form an unbiased estimator from it and show that new estimator is unbiased.

iii. If the MoM estimator of x_0 is unbiased, is it also consistent? If not unbiased, is the new unbiased form you made from it consistent? In either case, prove consistency or argue that the estimator is not consistent.

(Yes, this setup should remind you of one of your first exam problems.)