Stat/Math 430 – Mathematical Statistics
Final Exam – Take Home
Distributed May 6, 2015

Due May 14, 2015 by 3 pm to 306 SM

Instructions:

1. Show all work (on your own paper or via a compiled RMarkdown file). You may receive partial credit for partially completed problems, so you should try all parts of all problems. Turn in ONLY your FINAL solutions please. **Please write solutions legibly!**
2. The exam is open book and open notes. You may also refer to your homework, homework solutions, handouts, and your exams, as well as any other items posted on our course Moodle site.
3. You may not use any external references/texts and no online media (i.e. closed general internet).
4. You may use a calculator, Excel, R/RMarkdown, or any other statistical computing program to help with calculations. Accessing related help files (for example *?pnorm*) is permitted. A template is provided for RMarkdown to assist with data entry.
5. You may not discuss the exam with anyone but Prof. Wagaman until after the exam has been turned in by all students. If you ask me a question which warrants clarification, 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: Posted on Moodle. 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	21	13	16	16	13	8	13	100

1. Estimation and The Power of Simulation (21) – Adapted from Rice

Type	Count	Probability	RV
Starchy green	1997	$.25(2+\theta)$	X_1
Starchy white	906	$.25(1-\theta)$	X_2
Sugary green	904	$.25(1-\theta)$	X_3
Sugary white	32	$.25(\theta)$	X_4

Consider the following data for two categorical variables related to theoretical probabilities about genetics of plant offspring from self-fertilized heterozygotes (this means AB x AB as described in class). The two variables are content (starch/sugar) and color (green/white) from which four types of plants arise. The parameter θ is bounded between 0 and 1 and is related to the linkage of the underlying genes. The data is taken from Fisher 1958.

a. We want to test whether or not the data is consistent with the stated theoretical probabilities. Derive the MLE for θ, denoted $\hat{\theta}$, and find its value based on the data.
b. Perform an appropriate test procedure to determine if the data is consistent with the stated theoretical probabilities at a significance level of 0.05. Be sure to show work to validate your steps, and report your final real-world conclusion.
Now suppose we are really interested in θ due to its connection to linkage properties. In fact, we want to form a confidence interval for θ. Assuming our large-sample theory applies to $\hat{\theta}$, as an MLE, then we need to derive or estimate its variance in order to construct the confidence interval for θ. The derivation is involved, so perhaps we could use simulation to estimate its variance.
c. Explain how you would use a simulation to estimate the variance of $\hat{\theta}$. Be sure you address how randomization would be employed and state the necessary inputs for the simulation to run. (Think carefully about this!) I strongly suggest you write pseudocode to show what steps you would use in your simulation.
As you probably suspect, working with $\hat{\theta}$ is involved, so we could consider alternate estimators of θ.
d. If we used only X_1, the random variable associated with the count in the first cell, to come up with a natural estimate of θ, what would your estimator be? Denote this estimator $\tilde{\theta}$.
e. Evaluate the estimator $\tilde{\theta}$ on this data set. How does this estimator compare to $\hat{\theta}$, based on this data?
f. Using properties of the marginal distribution of X_1, show that $\tilde{\theta}$ is unbiased for θ and find an expression for its variance. (Simulation is not needed here as an exact expression can be found).

2. A Little Bit More Estimation (13)
Suppose $X_1, X_2, ..., X_n$ is a random sample of *n* observations from a Pareto distribution with pdf in the general case of: $f(x) = \frac{\alpha\beta^\alpha}{x^{\alpha+1}}$ where $x \geq \beta > 0$, and $\alpha > 0$, and 0, otherwise. Assume that β is known. We will also assume that $\alpha > 1$ for our derivations, but that it is unknown.

a. Find the method of moments (MoM) estimate of α.
b. Now assume $\alpha > 1$ is known, but β is not (just for this part of the problem). Find the MoME for β. Is the MoME unbiased? Show your work. If it is not unbiased, find an unbiased version of it, if possible.
c. Find the MLE of α.
d. Identify a sufficient statistic for α, and justify your choice.
e. Is your MLE minimal sufficient? Explain in one sentence.

3. Quadratic Regression with Constraints (16)

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 remaining regression assumptions are met – homoscedasticity (constant variance), independence, and normality of the error terms.

a. Determine the relevant least squares 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, 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. 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 for their respective parameters.

e. Find the variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ 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	-5	-3	-2	-1	0	1	2	3	5
Y	77.70	31.85	6.13	-1.87	-8.22	10.23	8.13	48.07	116.25

Note: You may use R for basic computations here (for example to find a mean or SD), and indeed you can use R to CHECK your computations for part b. and c., but you must show work for deriving the estimates and performing the various computations for this data set.

4. Likelihood Ratio and UMP Tests (16)

Suppose $X_1, X_2, ..., X_n$ is a random sample from a Poisson(λ) distribution.

a. Find the likelihood ratio test for testing $H_0 : \lambda = \lambda_0$ vs. $H_A : \lambda > \lambda_0$. Be sure to clearly state your test statistic, the distribution of the test statistic, and the form of the rejection region for a given significance level. You should NOT need to use the asymptotic result to get a distribution for the test statistic (but do that if you can't get one another way).

b. Complete problem 10.111 in your textbook (page 555).

c. Argue that the likelihood ratio test you found in part a. is UMP using your results from b. (Hint: it may help to pick a specific value under the alternative).

5. General Testing (13)

Two city council members are interested in estimating the proportion of voters who plan to vote in favor of a second term for them. The **first** city council member obtains a random sample of 15 voters and finds that 13 say they plan to vote in favor of a second term for the council member. The **second** city council members hires a surveying firm, which obtains a random sample of 150 voters and finds that 62 plan to vote in favor of a second term for the council member.

a. For the first council member, is there significant evidence that the proportion of voters who plan to vote in favor of a second term is greater than 2/3? As part of your response, compute an appropriate p-value, and explain what it measures and means in layman's terms. Use a significance level of 0.05.

b. For the second council member, obtain a 90% confidence interval for the proportion of voters who plan to vote in favor of a second term for that council member. As part of your response, explain what the 90% confidence level refers to.

c. Ignoring the council members' re-election chances, which council member has more reliable results based on the data collected and analysis above? Why? Explain.

6. Bayesian Inference (8) – adapted from Chihara and Hesterberg

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Uniform(0, θ) continuous distribution. A prior density for θ is selected to be the Pareto distribution, giving the prior the following form:

$g(\theta) = \frac{\alpha\beta^{\alpha}}{\theta^{\alpha+1}}$ where $\theta \geq \beta > 0$, and $\alpha > 0$, and 0, otherwise.

a. Explain why a normal prior on θ would not make sense in this context.

b. **Derive** the posterior distribution for θ based on the random sample of *n* observations. You may use a proportionality argument and drop constants, but need to show some work for obtaining the posterior distribution.

c. Assume that you observe the following four observations from the Uniform(0, θ) distribution: 6, 6, 8, 10, and that the Pareto prior on θ has parameters $\alpha = 0.3$ and $\beta = 5$. Fully specify the posterior distribution for θ.

d. State the formula for the Bayes estimate of θ and find its value in the same setting as part c.

7. Assessing the Fit (13)- data taken from Rice

A study examined birds which were feeding and researchers counted the number of hops between flights for each bird. The frequency of each count of number of hops is reported in the table shown:

# of Hops	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	48	31	20	9	6	5	4	2	1	1	2	1

We want to fit an appropriate model to this data.

a. Looking closely at the data, explain why a Poisson model would not be appropriate to fit to this data.

b. Find the MLE for p when fitting a geometric distribution to this data. In other words, fit a geometric distribution to the data, and report the "best-fitting" p. Note: in your textbook, the geometric is defined in terms of the trial number of the (first) success, where p is the probability of success.

c. Assess the fit from part b. using an appropriate GRAPH. Be sure to comment on the fit and explain your choice of graph.

d. Assess the fit from part b. using an appropriate inference procedure (with significance level 0.10). Be sure to check all relevant conditions, show justification for your work, and report a final conclusion.