Mathematics 376 – Mathematical Statistics
Review Sheet, Final Exam
May 7, 2012

General Information

The final examination for this course will be given at 11:30 a.m. on Wednesday, May 16 in our regular class room, Swords 302. The exam will be roughly one and a half times the length of one of the midterms. You will have the full two and one half hour period from 11:30 am to 2:00 pm to work on it if you need that much time. *Important Note: Because others may need the room and because some of you may have exams after this one, I will not be able to give you extra time this time.* As was true for the midterms, I will let you bring an 8.5 by 11 sheet of paper to the exam containing any information you want and you may consult it at any time. Copies of any necessary tables will also be provided. Even more so than on the midterms, of course, it will be necessary to prepare carefully; you don't want to have to search for every formula you need or a similar examples to help you get started on every problem. You will be pressed for time to complete the exam if you are doing that.

Topics to be Covered

1) Sampling distributions related to the normal distribution (χ^2, t, F distributions) – know how to tell when a random variable has one of these distributions, how to use the tables for each in the text, etc.
2) Point estimators for distribution parameters, bias, mean square error, standard error. Be sure you understand Table 8.1 on page 397 of the text, where all the entries come from, and how they are used. *Be aware that this table covers only the large sample case. The confidence intervals and tests for means and difference of means in the small sample case are slightly different because of the hypotheses needed to assure the relevant statistic has a t-distribution. Review Section 8.8 if this is not clear.* Also be able to analyze estimators to determine whether they are biased or not, construct unbiased estimators, etc.
3) The pdf's for order statistics (especially the sample maximum and minimum), and how they can be used for estimation problems, especially in conjunction with:
4) The method of moments and the method of maximum likelihood for deriving estimators. (The other material we discussed in Chapter 9 on consistency of estimators, sufficient statistics, etc. will *not* appear on this exam.)
5) Hypothesis testing – the general concepts: null hypothesis, alternative hypothesis, test statistic, rejection region, Type I error probability ($= \alpha$, or level of test), Type II error probability ($= \beta$), attained significance level (p-value of a test), interpretation of results.
6) The connection between confidence intervals and rejection/"acceptance" regions for tests.
7) Large sample (z-) tests and related confidence intervals for means and proportions. Questions here might also ask you to design tests with a given α-value to achieve a certain β-value by selecting sample size appropriately.

8) Small sample (t-) tests for means and related confidence intervals.
9) χ^2-tests for variances and related confidence intervals; F-tests for ratios of variances and related confidence intervals.
10) Least squares estimation for linear models of the form $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$. The kinds of questions I might ask here will be of two types:
    a) Be prepared to compute the least squares estimators for the coefficients β_i in a "small" simple regression ($k = 1$) example (say with $n \leq 10$ or so) with a calculator (`R` will not be available!). Also know how to carry out hypothesis tests on the β_i in this case.
    b) I might also ask you to set up the normal equations for the least squares estimators in a multiple regression ($k \geq 2$) example or answer some general questions about hypothesis testing in that situation. I will not ask any computational questions there, however.

Suggestions on How to Study

Start by reading the above list of topics carefully. If there are terms there that are unfamiliar or for which you cannot give the precise definition review them. Reread the class notes. *Everything on the final will be similar to something we have discussed at some point this semester.* Also look back over your graded problem sets and exams. If there are problems that you did not get the first time around, try them again now. Then go through the suggested problems from the review sheets. If you have worked these out previously, it is not necessary to do them all again. But try a representative sample "from scratch". Don't just look over your old solutions and nod your head if it looks familiar. Practice thinking through the logic of how the solution is derived again.

Suggested Review Problems

Look at the problems from the two previous review sheets for the topics 1 - 9 above. For the last one (regression and hyptheses concerning the regression coefficients): From Chapter 11/31, 63 ("linearize the model" means to take the logarithm of both sides of $E(Y) = \alpha_0 e^{-\alpha_1 x}$ as $\ln(E(Y)) = \ln(\alpha_0) - \alpha_1 x$ and do the regression with the data points $(x_i, \ln(y_i))$ to estimate $\beta_0 = \ln(\alpha_0)$ and $\beta_1 = -\alpha_1$), 69, 72 (Exam question could be "How would you set up a test of $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$?")

Review Session

If there is interest, I would be happy to run a review session after the end of classes. We can discuss this in our last class meeting.

Sample Exam Questions

General Notes:
(1) A "random sample" always consists of independent measurements from an indicated distribution.

(2) This is the exam from the Spring 2010 offering of MATH 376; this year's exam will be a bit shorter because the exam period is now only 2.5 hours long.

I. Let Y_1, Y_2, Y_3, Y_4, Y_5 be a random sample from a normal distribution with mean $\mu = 4$ and standard deviation $\sigma = 10$.

A) (10) What is the distribution of $\overline{Y} = \frac{1}{5}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5)$? Why?
B) (10) What is the distribution of $U = \frac{(Y_1-4)^2+(Y_2-4)^2+(Y_3-4)^2}{100}$? Why?
C) (10) What is the distribution of $V = \frac{\sqrt{3}(Y_4-4)}{10\sqrt{U}}$, where U is as in part B? Why?
D) (10) How would you determine the PDF for the sample maximum $Y_{(5)}$? (Note: The CDF for a normal random variable is not an elementary function; just give a "recipe" for how it might be computed.)

II. A random variable Y is said to have a *log-normal* distribution with parameters μ, σ if its pdf has the form

$$f(y) = \begin{cases} \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-(\ln(y)-\mu)^2/(2\sigma^2)} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

A) (10) Compute $E(\ln(Y))$ for a log-normal random variable. (Hint: Set up the integral, then change variables.)
B) (10) Let $Y_1, \ldots, Y_n$ be a random sample from a log-normal distribution with unknown μ and known $\sigma = 1$. Find the maximum-likelihood estimator for μ.
C) (5) Is your estimator from part B biased or unbiased? Why?

III. (15) Solid copper produced by melting powdered ore is tested for "porosity" (the volume fraction due to air bubbles). A sample of $n_1 = 40$ porosity measurements made in one lab has $\overline{y}_1 = .25$ and $s_1^2 = .001$. A second set of $n_2 = 50$ measurements is made using identical ore in a second lab, yielding $\overline{y}_2 = .17$ and $s_2^2 = .002$. Find a 95% confidence interval for $\mu_1 - \mu_2$, the difference of the population mean porosity measurements from the two labs. What conclusion can you draw from your interval?

IV. The Rockwell hardness index for steel is determined by pressing a diamond point into the metal with a specified force and measuring the depth of penetration. Out of a random sample of 50 ingots of a certain grade of steel from one manufacturer, the Rockwell hardness index was greater than 62 in 31 of the ingots.

A) (15) The manufacturer claims that at least 75% of the ingots of this type of it produces steel will have Rockwell hardness index greater than 62. Is there sufficient evidence to refute this claim? Use a test at the $\alpha = .01$ level.
B) (15) Using the rejection region you found in part A, compute the Type II error probability β of your test if it is actually true that 65% of the ingots have hardness index greater than 62.

V. Consider the following measurements of the weights of yields of two breeds of apple trees (in kilograms):

Breed 1 :	80.6	80.3	81.5	80.7	80.4	
Breed 2 :	79.5	79.9	81.0	79.4	79.2	81.4

Assume the measurements come from normal populations.

A) (20) Estimate the *variances* σ_1^2 and σ_2^2 of the yields of the two breeds. Is there evidence to suspect that $\sigma_2^2 > \sigma_1^2$? Explain.
B) (15) Test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the alternative $H_a : \mu_1 \neq \mu_2$. Estimate the p-value of your test and state your conclusion clearly and succinctly.

VI. The following table gives measurements of the amount of sodium chloride that dissolved in 100 grams of water at various temperatures in a chemistry experiment.

x (degrees C)	y (grams)
0	7.3
15	13.0
30	23.3
45	30.7
60	39.7
75	47.7

A) (20) Find the equation of the least squares regression line for this data set.
B) (15) Is there sufficient evidence to say that $\beta_1 > .45$? Explain, using the p-value of an appropriate test.

VII. (20) (A "Thought Question") Does a "statistically significant" result where we reject some H_0 mean that H_0 is *far from being true*? Answer intuitively first. Then answer the following: Suppose each individual we draw from a population has either property A or property B (but not both). Let p be the proportion of the population that has property A. We test $H_0 : p = 1/2$ versus $H_a : p > 1/2$ with large-sample tests with $n = 100, 1000, 10000, 100000$. What must the observed proportion of sampled individuals with property A be in order to reject H_0 at the $\alpha = .05$ level in each case? How does this square with what you said at first?

Extra Credit (20) Recall that in the "Big Theorem" in the multiple regression case, we said that in the entries $c_{ij}\sigma^2$ of the covariance matrix of the least squares estimators, the c_{ij} were the entries of the matrix $(X^tX)^{-1}$. Show this is true by direct computation for the X from a simple linear model of the form $Y = \beta_0 + \beta_1 x + \epsilon$.