Stat 135, Fall 2006 A. Adhikari

HOMEWORK 7 (due Friday 10/27)

1. Read the page `help(chisq.test)` all the way down. There will be stuff you don't understand, but most of it should make sense.

2. 9.34.

3. 9.38.

4. (From Freedman, Pisani, and Purves) In the Current Population Survey of March 2005, men were classified by employment status and marital status. Here are the data; you may assume they come from a simple random sample of U.S. men. "Once married" means, "widowed, divorced, or separated."

	married	once married	never married
employed	790	98	209
unemployed	56	11	27
not in labor force	21	7	13

a) Were marital status and employment status independent for U.S. men in 2005? State appropriate null and alternative hypotheses and use R to perform an appropriate test. Check that the degrees of freedom used by R are what you think they ought to be. (Ignore the warning message for now.) If you conclude that the variables were not independent, suggest some ways in which they appear to be dependent.

b) Print out what R calls the "expected counts under the null hypothesis." (**Don't** program R to compute the expected counts; learn how to get hidden output of R programs. It may help you do the simulation in Problem 10.)

Explain how R got the expected count in the "married and employed" cell, and use the null hypothesis to justify the calculation. Also explain why R gave a warning message.

5. In the test in the previous problem, is R computing $\sum(O_i - E_i)^2/E_i$, or $-2\log\Lambda = 2\sum O_i \log(O_i/E_i)$? Check by calculation.

6. Use the data in Problem 4 to construct an approximate 95% confidence interval for the proportion of unemployed men among all U.S. men in 2005.

7. If possible, use the data in Problem 4 to construct an approximate 95% confidence interval for the difference between the proportions of employed and unemployed men among all men in the U.S. in 2005.

8. 9.40

9. Generate an i.i.d. sample of size 1000 from the binomial distribution with parameters $n = 5$ and $p = 0.4$. Compute the counts in the categories 0, 1, 2, 3, 4, and 5. You shouldn't have to get each count separately. Think of a familiar program which does the counting as part of its calculations, and get output from that program.

Now fit the binomial model to your sample. That is, pretend you didn't know that p was 0.4, and fit the best binomial you can from your data. Do the χ^2 test. Explain the degrees of freedom that should be used, and explain any discrepancy in what you see in your R output. Get the p-value and state the conclusion of your test.

10. Now repeat the sampling in the previous problem 2000 times. That is, generate 2000 independent samples, each of size 1000, from the binomial distribution with $n = 5$ and $p = 0.4$. For each of your samples, get the counts in each category 0, 1, 2, 3, 4, and 5.

Now fit the best binomial to each of your samples, pretending that you didn't know that $p = 0.4$. For each sample, compute $X = -2\log\Lambda = 2\sum O_i \log(O_i/E_i)$ as well as its approximation $Y = \sum(O_i - E_i)^2/E_i$.

Draw the empirical (i.e. observed) histograms of X and Y. Superimpose the appropriate χ^2 curve over each histogram. Do your pictures agree with the theory indicated on pages 341-343 of the text?