**Stat 135, Fall 2006 A. Adhikari**

# HOMEWORK 5 (due Friday 10/6)

**1.** 8.20. Use $R$ to do this one accurately. Be careful about constants when you're working out the distribution of $\hat{\sigma}^2$.

**2.** 8.26. Is your estimate a method of moments estimate? Is it a maximum likelihood estimate?

**3.** 8.30.

**4.** 8.32. You don't have to do all six intervals in parts (b) and (c); just do the 90% intervals. If I'm feeling kind I'll type the data in for you; check the homework page on the web.

**5.** 8.48. This uses the $\delta$-method we discussed a couple of weeks ago.

**6.** There's nothing to turn in for this one, but I won't help you with figuring out how to get $R$ to do things if the methods are easily found by doing this exercise, or by using commands from old homeworks, or by methods outlined on our pages of tips.

Browse *An Introduction to R* and the $R$ reference card. Experiment with some of the commands. Also learn to use `help.search`. For example, if you want to know whether $R$ has a command that computes the gamma function (that's $\Gamma(\alpha)$ for positive $\alpha$), do:

> help.search("gamma")

In the left column you'll see a list of related commands, and, in parentheses, the packages in which they reside. The descriptions are in the right column. Clearly the one you want is `Special Functions of Mathematics`. So use the left column of that line as follows:

> help(Special, package=base)

You should get a useful help page.

The dataset `data.scores` contains the midterm (second column) and final (first column) scores of my Stat 2 class from last Spring. You'll find it on the homework webpage. All the problems below refer to that dataset.

**7.** Read the data into $R$. **Do not** type the numbers in; figure out how to get $R$ to read the matrix!

Clean up the data: keep only the rows where both entries are positive. (The other rows correspond to absences, students who dropped the class, etc.) The midterm was out of 20 and the final was out of 40. Multiply all the midterm scores by 2 so that they are on the same scale as the final. Switch the columns to put them in chronological order: midterm scores in Col 1, final scores in Col 2. **All subsequent exercises refer to this cleaned-up dataset.**

**a)** Draw the histograms of the two variables, making sure that the intervals are the same for both histograms to make direct comparison easier.

**b)** Read Section 6 of Chapter 10. A boxplot gives a quick visual sense of the skewness of a distribution, and also allows a quick comparison of several distributions. On a single graph, draw boxplots of the midterm scores and final scores. Comment on what the boxplot tells you about the center, spread, and skewness of the distributions, and compare with the histograms in part (a).

Also, look at the cover of your text. You now know what the picture is.

**8.** If a variable $X$ has a normal distribution, then it can be written as $X = aZ + b$ where $Z$ is standard normal. So if you plot the percentiles of $X$ against the percentiles of $Z$, you should get a straight line. This is the idea behind the main method in this problem.

**b)** Generate an i.i.d. sample of size 500 from the normal distribution with mean 10 and SD 3. Use `qqnorm` to plot the quantiles of your sample against the quantiles of the standard normal. This is called a "normal quantile plot". Does the plot look linear?

**a)** Now make the normal quanitle plot of the final scores and comment on linearity. Compare with the histogram in Problem 7.

**9.** John Tukey was an astonishingly clever man and came up with lots of quick "back of the envelope" methods for summarizing data. The boxplot is one such method. Here's another - it's called the stemleaf diagram, or stem-and-leaf plot. Your text has an example which starts at the bottom of Page 391 (Section 10.3). As the text says, the diagram is easier to describe by examples than by general definitions.

First run your eye over all the final scores (the numbers), and then look at the histogram you drew in Problem 7.

Now use the command `stem` on the final exam scores. You should get something that looks like a histogram on its side. It's fine, but it looks a little odd because of how it has grouped the data. So play with the *scale* option to the command. Run it once with `scale=0.5` and then again with `scale=2`. Of the three stemleaf diagrams, which one do you think is the best summary of the data? Why?

**10.** My grading schemes always allow the final to kill the midterm. (Check the course description for Stat 135!) So in Stat 2, a student would gain from this scheme if his/her final exam score was more than twice the midterm score. Remember that you've already multiplied the midterm scores by 2.

**a)** Plot the final scores (vertical axis) versus the midterm scores (horizontal axis). Because you have already rescaled the midterm scores, both variables should be in the 0-40 range. Draw the $x = y$ line on the plot. By eyeballing the plot, say whether the percent who gained from the grading scheme was greater than 50%, equal to 50%, or less than 50%.

**b)** Use $R$ to count how many students gained from the grading scheme, and check that this number is consistent with your eyeballed estimate in (a).

**c)** The plot looks roughly linear (there is some curvature, but the linear approximation is not terrible). That means it looks like points clustered around a straight line. Suppose you were trying to draw that straight line, with the idea of using the line to estimate final exam scores based on midterm scores. Would you use the line you drew in part (a)? If not, would the line you use be flatter (smaller slope) or steeper than the line in (a)? Give an intuitive justification of your answer.

Pretty soon we're going to be doing linear regression, which is exactly what you're trying to do in this problem: find the best straight line to estimate $y$ based on $x$.