Biost 517: Applied Biostatistics I
Emerson, Fall 2005

Homework #2 Key
October 7, 2005

<u>**Written problems:**</u> To be handed in at the beginning of class on Wednesday, October 12, 2005.

> *On this (as all homeworks) unedited Stata output is* <u>***TOTALLY***</u> *unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*

1. The class web pages contain descriptions of two datasets
   - SEP.doc
   - dfmo.doc (consider the "long" format)

   a. For each of the described scientific questions, briefly characterize the type of statistical question to be answered. That is, using the classification presented in class, characterize the problem as clustering of cases, clustering of variables, quantifying distributions within groups, comparing distributions across groups, or prediction, identifying any variable whose distribution is of interest and any groups that might be being compared.

<u>**ANSWER:**</u>

For the SEP dataset, the key problem is to derive an interval prediction to be used as a normal range. Along the way, we will be interested in deciding whether we need to derive different ranges by sex, age, and height. This latter aspect takes on some element of comparing distributions across groups. However, the major issue is one of prediction. The outcome variable will be one of n35, p40, n50, or p60 (alone or in combination) as measured on the left, right, or an average. The groups that might be compared are groups defined by all combinations of sex, age, and height (perhaps after dividing age and height into intervals).

For the DFMO dataset, we are interested in comparing the distribution of putrescine, spermidine, or perhaps the spermidine:spermine ratio across groups defined by dose of DFMO.

   b. For each of the datasets, classify the available measurements with respect to the statistical role they might play in answering the scientific question. That is, using the classification presented in class, identify which variables might be outcome measurements, predictors of interest, subgroup identifiers for interactions, potential confounders, precision variables, surrogates for the response, or irrelevant.

<u>**ANSWER:**</u>

For the SEP dataset, the n35, p40, n50, and p60 measurements are the outcome variable. Age, height, and sex are the grouping variables of greatest interest. Equipment and date would take on the role of potential confounders and/or effect modifiers (though we certainly hope not, in which case they would become irrelevant).

For the DFMO dataset, the outcome variables are putrescine, spermidine, and perhaps the spermidine:spermine ratio (perhaps at specific times). The predictor of interest is dose of DFMO, and time is an effect modifier unless we consider the outcome measurements at each time separately (note that by

randomization, the distiribution of put, spd, and spm will be the same across dose groups at baseline, but the treatment might cause there to be differences at other times). Because we have a randomized clinical trial, we are less concerned with confounding by sex or age, though we could consider including them as precision variables if it were known that they were strong predictors of polyamine status. Spermine could either be viewed as part of the outcome (if we analyze the ratio) or as a precision variable that we would want to adjust for.

     c. For each of the datasets, classify the available measurements with respect to the type of measurement: qualitative versus quantitative, unordered versus partially ordered versus ordered, discrete versus continuous, and interval versus ratio.

ANSWER:

For the SEP dataset, the n35, p40, n50, and p60 measurements are continuous (ratio) variables, as are age and height. Sex is a binary random variable. As coded, date is merely a label, though we could convert it to a continuous measurement of days since some fixed date (this is termed a "Julian date"). Equipment is binary in our dataset, but would probably best be viewed as an unordered categorical variable, because there are probably other types of equipment. Patient ID is an unordered categorical variable.

For the DFMO dataset, the outcome variables putrescine, spermidine, and spermine are all continuous (ratio) variables, as is age. Dose and time are also continuous measurements, although they were sampled discretely. Sex is binary, and patient ID is an unordered categorical variable.

2. This problem deals with a data set hw2salary.txt which represents three samples of monthly salary (in US dollars) for male and female university faculty. For each sample (labeled "A", "B", "C"), salaries for 1000 men and 1000 women were sampled. Of particular interest is whether there is evidence that men tend to be paid more than women in the populations from which each of the samples were drawn.

    Using the three samples of salary data in the file hw2salary.txt, generate the following descriptive statistics for each sex in each sample.
- Histogram
- Number of cases with missing data
- Mean
- Geometric mean
- Median
- Mode (it suffices to take an approximate mode from a histogram)
- Standard deviation
- Variance
- Minimum and maximum
- Range (the difference between minimum and maximum)
- 25th, 75th percentiles
- Interquartile range (the difference between 25th and 75th percentiles)
- Proportion of subjects earning at least $50,000 per year
- Proportion of subjects earning at least $75,000 per year

    For each sample, how would you answer the question regarding whether men tend to be paid more than women?

ANSWER:

For each sample, descriptive statistics are presented for each sex. These descriptive statistics include a histogram, the mean, standard deviation, geometric mean, mode, minimum, 25th percentile, median, 50th percentile, maximum, range, interquartile range, percent earning more than $50,000 per year, and percent earning more than $75,000 per year.

Note that this problem merely asked you to decide whether the men were paid more than the women in the sample. There was no need to do inferential statistics, because we were not asking about some larger population. (Recall that the second question in my two question test of whether you need a statistician asks whether you know how to decide what is larger when you have all the pertinent data present. This problem is to show that this can be a hard task.)
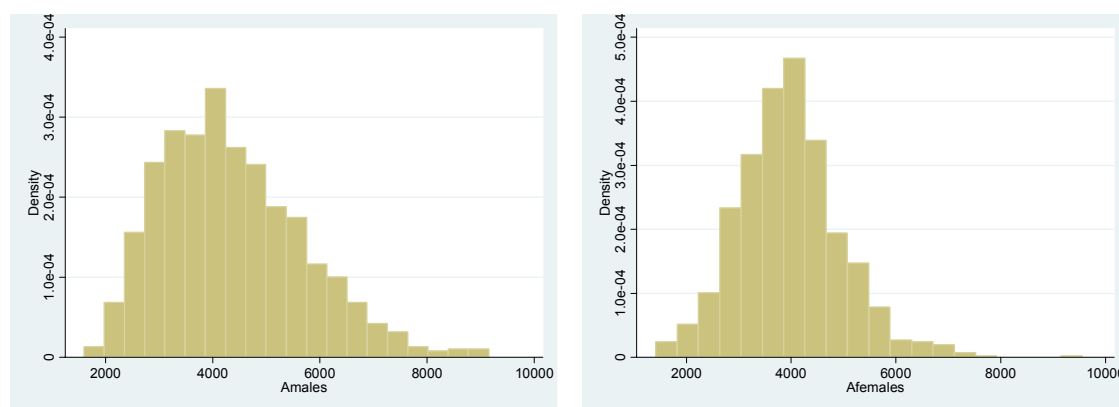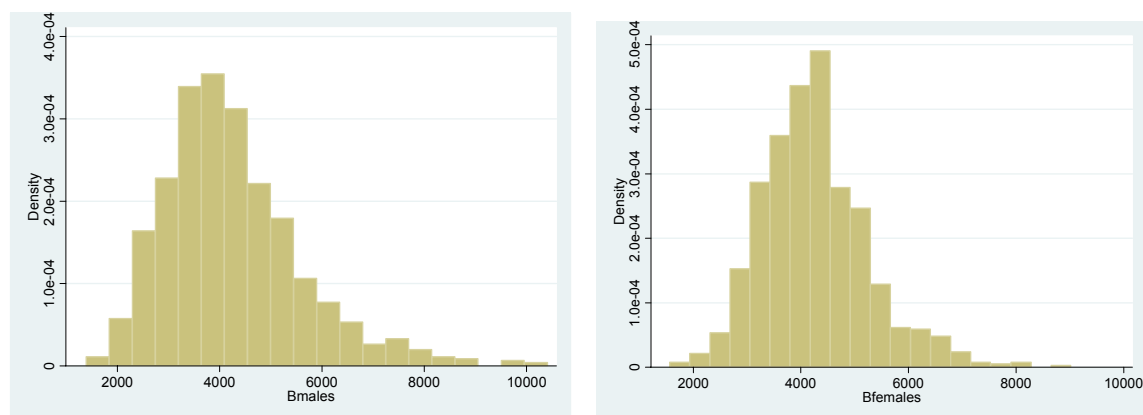
Sample A

Figure 1: Histograms of monthly salaries for males (left) and females (right) at University A.
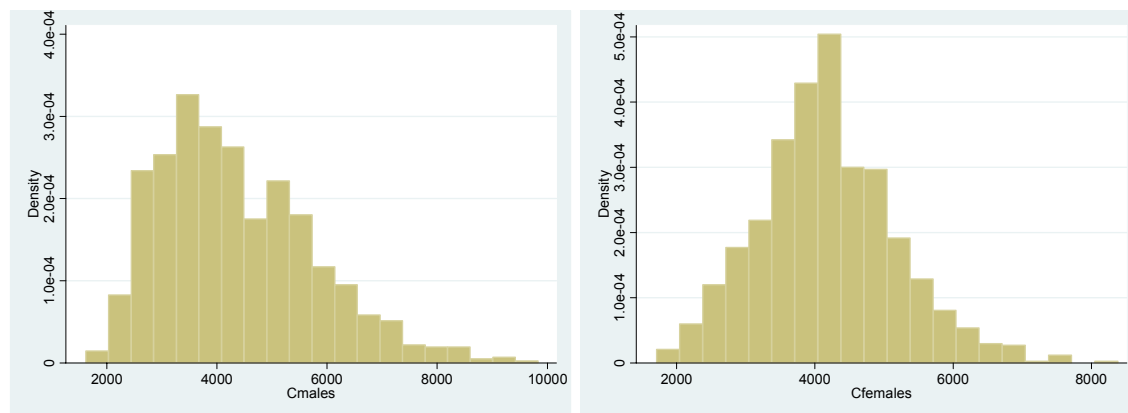
	Mean	Std. Dev.	Geom Mean	Mode	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k	Range	IQ Range
SAMPLE A													
Males	$4366	$1332	$4170	$3962	$1600	$3361	$4174	$5160	$9154	50.6%	9.1%	$7554	$1799
Females	$3968	$987	$3844	$4004	$1418	$3313	$3933	$4535	$9556	37.7%	2.4%	$8138	$1219

Ans: From the above statistics, the males would appear to tend to be paid more than females. This conclusion would be reached no matter whether we focus on the mean, geometric mean, median, or the percentage with salaries in excess of $50,000 or $75,000.

Sample B

Figure 2: Histograms of monthly salaries for males (left) and females (right) at University B.

	Mean	Std. Dev.	Geom Mean	Mode	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k	Range	IQ Range
SAMPLE B													
Males	$4247	$1348	$4050	$3850	$1392	$3296	$4048	$4951	$10411	45.6%	8.1%	$9019	$1654
Females	$4271	$993	$4159	$4251	$1563	$3607	$4220	$4812	$9025	52.1%	4.7%	$7462	$1204

Ans: Making a single decision about this sample is more difficult, and we would have to consider which measure is most important to us. From the above statistics, the females would appear to tend to be paid more than males when judging by the geometric mean, median, or the percentage with salaries in excess of $50,000, but more men appear to receive salaries in excess of $75,000. (I note that the difference in average salaries does not appear to be all that great.)

Sample C

Figure 3: Histograms of monthly salaries for males (left) and females (right) at University C.

	Mean	Std. Dev.	Geom Mean	Mode	Min	25th %ile	Median	75th %ile	Max	% > 50k	% > 75k	Range	IQ Range
SAMPLE C													
Males	$4362	$1399	$4150	$3653	$1623	$3298	$4105	$5242	$9833	47.8%	10.2%	$8210	$1940

Females	$4183	$1004	$4061	$4106	$1711	$3559	$4112	$4806	$8381	48.1%	3.1%	$6670	$1247

Ans: Making a single decision about this sample is also difficult, and we would have to consider which measure is most important to us. From the above statistics, the females would appear to tend to be paid more than males when judging by the 25th percentile (a rather unusual summary measure to use, but not necessarily inappropriate). There does not appear to be any scientifically meaningful difference in the median salaries or the percentage making more than $50,000. Men are on average paid more, however, in large part due to the fact that the percentage of men making more than $75,000 is markedly higher. The geometric mean is also higher for men.

   3. For the males at University A:
   a. Give one person (tell which case you use by row number) a raise (tell how much of a raise) that would increase the mean monthly salary for men at that university by at least $1,000 (tell what the resulting mean monthly salary is).

Ans: I can give any single person a raise of $1,000,000 per month, and that will raise the average salary by $1,000. I can give any single person a raise of $50,000 per month, and that will raise the average salary by $50. Basically, if I give a single person a raise of X dollars per month, that will raise the average salary by X/1000 dollars.

   b. Give one person (tell which case you use by row number) a raise (tell how much of a raise) that would increase the geometric mean monthly salary for men at that university by at least $1,000 (tell what the resulting mean monthly salary is).

Ans: This is a bit harder to figure out. I would have been perfectly happy with you doing it by trial and error. But we can figure it out on the logarithmic scale. The geometric mean was $4,170. This means that the average of the log salaries was log(4170) = 8.33567, and the total of the log salaries was therefore 8335.67. We want the geometric mean to be $5,170, which corresponds to an average of the log salaries to be log(5170) = 8.55063 and a total of the log salaries to be 8550.63. So we need to increase one person's log salary by 8550.63 – 8335.67 = 214.96. Of course, increasing the log salary by 214.96 is the same as multiplying that salary by $e^{214.96}$ = 2.27 x 10^{93}. If I want to save money doing this, I might as well give that raise to the male with the lowest salary. So the new salary would be $1600 x 2.27 x 10^{93}, a salary that I would not turn down. If all we need to do is raise the geometric mean salary by $50, the same sort of computation finds that a 1000 x log(4220/4170) = 150105.3 fold increase in some single person's salary will do the trick. Again, in order to save money, I might take the person earning $1600 per month and increase his salary to $240,168,480 per month.

   c. Explain why you cannot give one person a raise and increase the median monthly salary for men at that university by $1,000, What is the maximum increase in median monthly salary that you can attain by giving one person an arbitrarily large raise?

Ans: If I give a raise to the person having the median salary, the most I can do is make the person who was previously just above him be the new median. As luck would have it, there were three people tied at the median salary, and giving one of them the raise would still have another person with that same salary be the median.

   d. What does the above say about the influence that an outlier can have on the group mean, geometric mean, or median?

Ans: This was just an exercise in demonstrating how sensitive the mean, geometric mean, and the median were to outliers. The mean is much, much more sensitive than the geometric mean, and the median is relatively unaffected by outliers. Sensitivity to outliers can be either good or bad,

depending upon the scientific question. If you have a treatment directed at removing outliers (e.g., eliminating the few rapidly dividing cells in a person's body), the median cell division rate for a person might never find out that the treatment was working. On the other hand, playing the lottery might not look so bad on average, but the vast majority of people just lose their money.