Biost 517: Applied Biostatistics I
Emerson, Fall 2010

Homework #3 Key
October 31, 2010

Written problems: To be handed in at the beginning of class on Wednesday, October 20, 2010.

Problems make use of the university salary data (salary.txt). The class web pages contain an annotated Stata log file (initsalary.doc) illustrating the way in which this data can be input into Stata. In particular, I illustrate how string variables can be encoded and how labels can be associated with particular values of variables. Because this is a very large file, you may also have to tell Stata to increase the amount of memory it is using for data.

Questions for Biost 514 and Biost 517:

1. Generate appropriate descriptive statistics on all relevant variables for all measurements in the dataset according to sex.

Answer:

From the following tables of descriptive statistics, it can be seen that approximately 80.2% of the available measurements pertain to male faculty members. The measurements on females are more likely to represent faculty members with neither a Ph.D. nor a professional degree (14.5% for females, 6.8% for males) , more likely to represent faculty in the Arts (20.9% for females, 12.9% for males) and less likely to represent faculty in the Professional Schools (9.9% for females and 21.6% for males), more likely to represent assistant professors (37.2% for females and 16.3% for males), and less likely to represent years in which the faculty member had administrative duties (7.5% for females and 11.3% for males). Records pertaining to male faculty tended to correspond to faculty members who received their degrees and started at the university earlier, and thus had more records pertaining to earlier times in the period covered by these data. The average of the monthly salaries reported for records pertaining to male faculty members was higher than that for females ($4,855 vs $4,185).

	Males (n=15,866)	Females (n=3,926)	All (n=19,792)
Degree			
Other	1,071 (6.8%)	569 (14.5%)	1,640 (8.3%)
Ph.D.	13,586 (85.6%)	3,220 (82.0%)	16,806 (84.9%)
Professional	1,209 (7.6%)	137 (3.5%)	1,346 (6.8%)
Field			
Arts	2,038 (12.9%)	802 (20.4%)	2,804 (14.4%)
Other	10,408 (65.6%)	2,735 (69.7%)	13,143 (66.4%)
Professional	3,420 (21.6%)	389 (9.9%)	3,809 (19.3%)
Rank			
Assistant	2,588 (16.3%)	146.0 (37.2%)	4,048 (20.5%)
Associate	5,064 (31.9%)	1,465 (37.3%)	6,529 (33.0%)
Full	8,210 (51.8%)	1,001 (25.5%)	9,211 (46.4%)
Admin Duties			
None	14,080 (88.7%)	3,633 (92.5%)	17,713 (89.5%)
Some	1,786 (11.3%)	293 (7.5%)	2,079 (10.5%)

	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
				Males (n=15,866)			
Year	87.2	5.59	76	83	88	92	95
Year of Degree	71.0	8.17	48	66	70	76	96
Start Year	75.0	8.83	48	68	74	82	95
Monthly Salary	4,855	2,064	1,200	3,322	4,505	6,040	14,464
				Females (n=3,926)			
Year	88.5	5.28	76	85	90	93	95
Year of Degree	76.6	8.34	54	71	76	83	95
Start Year	80.4	8.09	57	74	80	87	95
Monthly Salary	4,185	1,523	1,267	3,152	3,996	4,951	11,036
				Both Sexes (n=19,792)			
Year	87.4	5.56	76	83	88	92	95
Year of Degree	72.1	8.50	48	67	72	78	96
Start Year	76.1	8.95	48	69	76	83	95
Monthly Salary	4,722	1,987	1,200	3,287	4,353	5,794	14,464

2. Suppose we want to compare the salaries paid to men to those paid to women over the time period covered by this data.

    a. From your results in problem 1, compare the mean of all salaries recorded for women in the dataset to the mean of all salaries recorded for men. (No statistical test need be performed. Just compare the descriptive statistics.) What scientific question would this be addressing?

Answer:

The average of the monthly salaries reported for records pertaining to male faculty members was higher than that for females ($4,855 vs $4,185). However, such an observation does not seem relevant to any important scientific question: There are variable numbers of observations on each subject, and the sampling scheme was such that an individual must have been employed by the university in 1995 in order to have any of their data represented. Because it appears that the women represented in the data set tended to be hired more recently, the data for men and women does not appear to be comparable when all records are included indiscriminately.

    b. Generate appropriate descriptive statistics for salaries paid to Associate Professors by sex. Compare the mean salary paid to female Associate Professors to that paid to male Associate Professors. What scientific question would this be addressing?

Answer:

The average of the monthly salaries reported for records pertaining to male faculty members while they were associate professors was lower than that for females ($3,966 vs $4,013). However, such a comparison is again problematic, and this analysis does not address any important scientific question due to unequal representation of males and females in the multiple measurements and because of likely confounding by economic trends over calendar time. That is, multiple measurements are made on the same faculty member, with likely differences between the sexes in the number of such measurements. Furthermore, as there was a tendency for males to have been hired earlier, the salaries

are not measured in a way that would account for inflation. Hence, the preponderance of males among the full professors, means that their salaries as associate professors would have been more likely to pertain to the late seventies (prior to high inflation rates) than would be true for females.

	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
				Males (n=5,064)			
Monthly Salary	3,966	1,391	1300	2,869	3,860	4,743	9,723
				Females (n=1,465)			
Monthly Salary	4,013	1,092	1,614	3,228	4,053	4,670	9,558
				Both Sexes (n=6,529)			
Monthly Salary	3,977	1,330	1,300	2,933	3,928	4,743	9,723

c. Compare the mean salary paid to women in 1995 to the mean salary paid to men in 1995. What scientific question would this be addressing?

Answer:

The average of the monthly salaries reported paid to male faculty members in 1995 was higher than that for females ($6,732 vs $5,397). Hence the average salary paid to men was 24.7% higher than the average paid to women (or, alternatively, the average salary paid to women was 19.8% lower than the average paid to men). This comparison can be used in part to address the question of discrimination against women by the university in the sense that if we assume that women are equally talented, trained, experienced, and assigned equal duties and if we assume the university does not discriminate in the rate of pay to women, we would expect the average salaries to be comparable for men and women in 1995 (note that we are now making comparisons on an individual basis—each individual is represented once). (I do note that even if these numbers were comparable, the university could still be discriminating in the number of women hired.)

	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
				Males (n=1,188)			
Monthly Salary	6,732	2,090	3,131	5,088	6,313	7,935	14,464
				Females (n=409)			
Monthly Salary	5,397	1,481	3,042	4,292	5,016	6,135	11,036
				Both Sexes (n=1,597)			
Monthly Salary	6,390	2,037	3,042	4,743	5,962	7,602	14,464

d. Compare the mean salary paid to female Associate Professors in 1995 to the mean salary paid to male Associate Professors in 1995. What scientific question would this be addressing?

Answer:

The average of the monthly salaries reported paid to male associate professors in 1995 was higher than that for females ($5,480 vs $5,019). Hence the average salary paid to male associate professors was 9.2% higher than the average paid to female associate professors (or, alternatively, the average salary paid to women was 8.4% lower than the average paid to men). This comparison can be used in part to address the question of discrimination against women by the university in the sense that if we assume that women are equally talented, trained, experienced, and assigned equal duties and if we

assume the university does not discriminate in the rate of pay to women when of comparable rank, we would expect the average salaries to be comparable for men and women associate professorsin 1995 (note that we are now making comparisons on an individual basis—each individual is represented once). (I do note that even if these numbers were comparable, the university could still be discriminating in the number of women hired *and* the number of women promoted.)

	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
				Males (n=299)			
Monthly Salary	5,480	1,224	3,769	4,604	5,095	6,223	9,723
				Females (n=138)			
Monthly Salary	5,019	858	3,880	4,416	4,743	5,318	9,558
				Both Sexes (n=437)			
Monthly Salary	5,335	1,141	3,769	4,538	5,022	5,962	9,723

e. How does the difference between the sexes computed in part d compare to the difference you found in parts a-c? Which is more appropriate to address the question of a systematic tendency to discriminate against women? Why?

Answer:

As noted above, the analyses in part a and b are not directly comparable to those in parts c and d due to the indiscriminate use of repeated measurements on the same individual in parts a and b.

In terms of assessing discrimination in the broadest sense (so historical as well as current discrimination), the analysis in part c addresses the most pertinent issues *except* the question of discrimination in hiring and firing. That is, the analysis in part c does not consider whether males and females are represented in equal numbers. However, it does presume that women are equally talented (I think this a safe assumption), equally trained (this was not historically true—there were not as many women in graduate school 40 years ago), equally experienced (again, historical discrimination comes into play), assigned equal duties (historical discrimination might mean that women are not sufficiently senior to be department chairs and women are not equally represented in senior positions in some fields), and women were promoted equally (but if there were evidence against this, I would consider that current discrimination).

By only comparing individuals of the same rank in part d, we are giving the university license to discriminate in salaries, so long as they make sure they do not promote the women. (This would tend to keep women in the lower paying ranks, but our analysis in part d would not consider the numbers of women versus the numbers of men in each of the ranks).

3. In problems 1 and 2, you generated descriptive statistics using all measurements in the dataset. However, multiple measurements were made on each subject. This problem guides you through the process of using Stata to determine how many repeat measurements are made on each individual.

   The data file contains repeated measurements on each individual. When our interest is on how individuals fare, we often combine such repeated measurements into a single

summary. For instance, we might consider taking the average of the measurements, the maximum or minimum of the measurements, or only the last measurement. Stata provides a command "egen" that will allow us to easily abstract such summaries by individual.

For instance, suppose we want the mean salary paid to each individual. We can obtain a variable *mnslry* that will contain that by:

- **`egen mnslry = mean(salary), by(id)`**

Each row will now have a value for variable *mnslry* that is equal to the mean of all the salary values for that individual. If you wanted to have instead the mean of salaries paid during the years 1990 - 1995 you could use:

- **`egen mnslry = mean(salary) if year > 89, by(id)`**

After this command, you would have a variable that had missing values for any rows corresponding to years 1989 or earlier, and for all other rows, the value for variable *mnslry* would be equal to the mean of all salaries paid in 1990 or later for that individual.

In this and the following problems you will need to use "egen" repeatedly in order to be able to perform analyses on a per individual rather than per measurement basis.

a. Use "egen" to generate a variable *nassoc* counting the number of years of available data for each individual as an Associate Professor, and provide suitable descriptive statistics for this variable using all cases in the datafile. The following Stata code will generate a variable *grbg*:

```
egen grbg= count(rank) if rank==2, by(id)
```

This variable will have missing data for any case in which *rank* was not 2 (i.e., not an Associate Professor). In order to obtain an entry for every row in the data set, we again use "egen":

```
egen nassoc= mean(grbg), by(id)
```

Even after this command, there will still be missing data for any individual who was never an Associate Professor. Hence, we can now enter 0 for all cases in which *nassoc* is missing.

```
replace nassoc= 0 if nassoc==.
```

How many measurements in the datafile correspond to a faculty member having a maximum of exactly 6 years as an Associate Professor? How many individuals does this represent? You might consider either or both of the following Stata commands:

```
table nassoc
list id rank nassoc if nassoc==6
```

Answer:

There were 1,518 records in the data file that corresponded to faculty members who had exactly 6 years worth of data as an associate professor. But there were only 103 faculty members who met these criteria (some of those faculty members had been hired as assistant professors, and some went on to get promoted and have some data pertaining to the years they were full professors.)

b. As can be seen in part a, doing descriptive statistics on the summarized variable is still complicated due to the number of repeated measurements on each individual. If we want to find out the distribution of *nassoc* across individuals (rather than rows in the file), we will need to restrict our analysis to one row for each faculty member. In this data set every subject should have had a 1995

measurement. We can check that by considering the maximum value of *year* for each individual. Generate a variable *maxyear* containing the last year for which a subject has a row in the data set, and provide summary statistics to show that each subject has a 1995 measurement. The following Stata code can be used to generate *maxyear*:

```
egen maxyear=max(year), by(id)
```

Answer:

All subjects did indeed have a record in 1995.

c. Now, since we know that every individual has a row corresponding to 1995, when we desire statistics on each individual, we could obtain summary statistics just for rows corresponding to *year==95*. Describe the distribution of the number of measurements made on each subject. Provide descriptive statistics that allow us to compare the number of measurements per individual by sex. What might be the scientific importance of any differences between the sexes? What might be the statistical ramifications of any differences? Are there differences that concern you?

Answer:

The following table presents descriptive statistics for the number of records per individual faculty member. Readily seen is the higher average number of records for males (13.4) versus that for females (9.6). This would mean that women would tend to have less seniority (thereby tending toward lower salaries), but also that the data for women would tend to be from more recent years (thereby tending toward higher salaries due to inflation). It also means that men will tend to be overrepresented in any analysis that does not take into account the multiple measurements made on each individual.

	N	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Male	1188	13.4	6.57	1	7	15	20	20
Female	409	9.6	6.34	1	4	8	16	20
Total	1,597	12.4	6.72	1	6	13	20	20

4. Generate a variable *massoc* reflecting the average of all salary measurements made for each individual while an Associate Professor.
   a. Provide summary statistics for *massoc* for the two sexes using all available data in the data set. What scientific question could be addressed using these descriptive statistics?

Answer:

The following table presents descriptive statistics for the average salary paid to each individual as an associate professor as recorded in each record. This analysis is of no scientific interest, because it includes variable numbers of records on each individual. In particular, men are overrepresented in this analysis.

	N	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Male	11,100	3,664	1,253	1,839	2,622	3,449	4,468	8,708

Female	3,095	3,851	984	1,808	3,133	3,978	4,379	8,836
Total	14,195	3,704	1,202	1,808	2,735	3,592	4,409	8,836

b. Provide summary statistics for *massoc* for the two sexes when each faculty member is represented only once. What scientific question could be addressed using these descriptive statistics?

Answer:

The following table presents descriptive statistics for average salary paid to each faculty member over the years he/she was an associate professor at the university. Faculty members employed in 1995 are not represented in this analysis if they were never an associate professor at the university between years 1976-1995. While this analysis treats individuals by appropriately including only one measurement per individual, the fact that variable numbers of records went into computing the means for each individual means that the measurements are still not particularly comparable: they are made over different time frames and are thus influenced by inflation.

	N	Mean	Std Dev	Min	25th %ile	Median	75th %ile	Max
Male	746	4,001	1,351	1,839	2,887	3,968	4,812	8,708
Female	238	4,125	1,020	1,808	3,609	4,165	4,605	8,836
Total	984	4,031	1,280	1,808	3,032	4,044	4,740	8,836

c. How do the standard deviation computed in part b differ from the standard deviations computed in problem 2b? Does this make sense? Why?

Answer:

The standard deviations were larger in problem 2b, because each record for an individual had different salaries. Thus there was more variability in that problem than in this problem, where we reduced the measurement for each individual to a single number.

d. Using the results for part b, do you worry about outliers in the data? Explain.

Answer:

In comparing the mean to the median, the mean is only slightly higher than the mean. Furthermore, the median is approximately midway between the 25^{th} and 75^{th} percentiles. So the bulk of the data seems fairly symmetric. However, the maximum is further from the median than is the minimum, so there may be some slight skewness to the data. It does not appear that any extreme outliers exist, however.

5. Comparisons made in problem 4 controlled for the rank of the faculty member. Why might controlling for rank be inappropriate when looking for salary discrimination by sex?

Answer:

Rank is intricately related to salary. Hence, controlling for rank is allowing the university to discriminate in salaries, providing they just refuse to promote women. I tend to think the best analysis would not adjust for rank, though we might want to do a

secondary analysis adjusting for rank to be able to describe mechanisms of discrimination. We might say that women are paid 6% less than a comparable man, some of which arises through 4% lower pay when in the same rank, with the remainder due to discrimination in promotion.

6. The comparisons of salary across sex groups might be potentially confounded by other variables.
    a. Would you *a priori* (before looking at the data) suspect that comparisons of salary might be confounded by academic field? Explain. Provide descriptive statistics that explore the possibility of such confounding.

Answer:

Supply and demand would likely mean that field affects the salary paid. It is also quite plausible that women have a harder time getting into some fields than others (I knew several women medical students who were told that women did not belong in surgery, for instance). Now the question is whether this last aspect is in the causal pathway of interest. To the extent that the university cannot now address historical trends toward women not entering into certain fields, we might choose to consider field as a confounder. Similarly, if women just do not like high paying fields like medicine, engineering, physics, etc. and prefer music, sculpture, art, etc., then we would want to adjust for field.

To examine evidence for these associations in the data, we note that in 1995, 19.56% of women were in the Arts, while only 11.8% of men faculty were in the Arts. Examined within the sexes separately, Faculty in the Arts were paid less than other Fields: $5,488 average monthly salary for men in the Arts vs $6,642 or $7,643 for men in the Other or Professional fields, respectively. (Similar patterns exist for females, though we sometimes just assess the association in one level of our predictor of interest.)

    b. Would you *a priori* (before looking at the data) suspect that comparisons of salary might be confounded by calendar year? Explain. Provide descriptive statistics that explore the possibility of such confounding.

Answer:

Inflation would likely mean that calendar year affects the salary paid. It is also quite plausible that women were discriminated against (or voluntarily chose not to work) historically. To the extent that the university cannot now address historical trends toward women not being hired long ago, we might choose to consider calendar year as a confounder.

In order to avoid using data on the same subject more than once, and also to compare the treatment of women at comparable stages in their career, we might choose to examine how starting salaries in the year of obtaining their degree are associated with calendar year. When we examine the association between sex and year of degree being the year hired, we see that newly hired women tended to average 1981 as their year of degree, while newly hired men averaged 1974 as their year of degree. Then among such faculty, those men who received their degree and were hired in 1980, 1985, 1990, or 1995 averaged starting monthly salaries of $2,093, $3,230, $3,873, and $4,278, respectively. Hence, there does appear to be a definite association between calendar year and salary.