Biost 517: Applied Biostatistics I
Emerson, Fall 2005

Homework #3 Key
October 21, 2005

Written problems: To be handed in at the beginning of class on Wednesday, October 19, 2005.

> *On this (as all homeworks) unedited Stata output is **TOTALLY** 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.)*

The following 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 will also have to tell Stata to increase the amount of memory it is using for data. All of the following variable names refer to the definitions in that file.

1. Using the description of the scientific question posed in the documentation for this data, briefly characterize

    a. The overall goal of the study.

    Ans: The overall goal is to see whether the salary of a female faculty member is lower because of her sex. Ostensibly, we would like to know the salary that would have been given to a male who was equally qualified in all respects.

    b. The specific aims of the data analysis.

    Ans: The specific aim is to use the data at hand to compare the salaries of females at the university to males who are most comparable. I am most interested in salary discrepancies that might be attributable to current actions at the university, rather than differences that result from historical discrimination. Because some characteristics of the faculty position might reflect aspects of both current and historical discrimination, the comparisons will need to be made in several stages.

    c. The scientific role of each of the variables.

    Ans: Monthly salary in 1995 is the measure of compensation, and faculty sex will be the predictor of interest. Variables measuring highest degree obtained, years since obtaining that degree, starting year at the university, and field represent some of the factors that influence salaries, but might also be considered as variables in the causal pathway of discrimination, depending upon the focus on current vs past discrimination and the possibility of discrimination in hiring practices. Rank is almost certainly in the causal pathway of interest, as it is intimately associated with salary.

    d. The statistical role of each of the variables.

    Ans: Monthly salary in 1995 will be the primary response variable, and faculty sex will be the predictor of interest. Variables measuring highest degree obtained, years since obtaining that degree, starting year at the university, and field are potential confounders or mediators of discrimination, depending upon the focus of a specific analysis (I will likely do several analyses and highlight the

differences in their interpretation). Rank will be a variable that I would include in the model only to describe a mechanism for discrimination.

2. Provide a table of sample size, number of missing observations, means, standard deviations, medians, quartiles, minima, and maxima for all of the variables in the dataset.

Variable	N Msng	Obs	Mean	Std. Dev.	Min	25th %ile	Median	75th %ile	Max
Case	0	19792	9896	5714	1	4948	9897	14845	19792
Faculty ID	0	19792	883	506	1	461	873	1315	1770
Year of Degree (19xx)	0	19792	72.1	8.50	48	67	72	78	96
Year of First Hire (19xx)	0	19792	76.1	8.95	48	69	76	83	95
Year (19xx)	0	19792	87.4	5.56	76	83	88	92	95
Administrative Duties (0= No, 1= Yes)	0	19792	0.11	0.31	0	0	0	0	1
Monthly Salary (dollars)	0	19792	4722	1987	1200	3287	4353	5794	14464
Female (0= M, 1= F)	0	19792	0.20	0.40	0	0	0	0	1
Degree	0	19792	1.99	0.39	1	2	2	2	3
Field	0	19792	2.05	0.58	1	2	2	2	3
Rank (1= Assist, 2= Assoc, 3= Full)	4	19788	2.26	0.78	1	2	2	3	3

a. For each variable, indicate the type of measurement represented by that variable (binary, unordered categorical, ordered categorical, discrete quantitative, continuous quantitative, censored) and identify the descriptive statistics which have no value to answer any scientific question related to that type of variable.

Ans: The number of missing and nonmissing observations are of course pertinent for all variables. Year of highest degree, year of first hire, year, and salary are quantitative continuous variables, and thus all the descriptive statistics are of some value. The indicators of administrative duties and females sex are binary variables coded as 0-1. Thus the mean is the proportion of faculty who have administrative duties or are female, respectively. The remainder of the descriptive statistics are not invalid, but are certainly boring. Rank is an ordered categorical variable. The mean and standard deviation are of very limited use on their own (but can provide some information when comparing two groups). All of the quantiles are of course valid for any ordered variable, though with only three distinct levels, the quantiles are not particularly informative. Degree, field, case, and id are all unordered categorical variables. None of the descriptive statistics are of any use for these variables.

b. Based on the descriptive statistics you obtained above, are there any outliers that you would worry about for any of the variables?

Ans: Salary has a mean larger than the median, and the median is not the midpoint of the range, nor is it midway in the interquartile range. This certainly suggests a skewed distribution, and there may well be some outliers. This is also suggested by the fact that the minimum is 1.5 SD below the median, while the maximum is 5 SD above the median. (It would take looking at the distribution of the data in a density estimate, histogram, or boxplot to really judge whether I thought there were serious outliers.) None of the other quantitative variables show signs of outliers.

c. If your interest is describing such quantities as the proportion of men and women employed at the university, the distribution of ranks, the distribution across fields, the distribution of salaries, etc., of what scientific value are the descriptive statistics you obtained in this problem? Briefly explain your answer.

Ans: The above descriptive statistics are based on multiple measurements for some individuals (indeed, any faculty member who was hired prior to 1995). Because the number of measurements per individual is not balanced across individuals, the above descriptive statistics will not be representative of the distribution of sex, field, etc. in the faculty at the university. We would rather have descriptive statistics on a sample where each individual is represented once.

3. Provide suitable statistics to describe the data pertinent to the faculty employed in the year 1995.

Variable	Obs	Mean	Std. Dev.	Min	25th%ile	Median	75th%ile	Max
Year of Highest Degree	1597	76.09	9.86	48	69	76	84	96
Starting Year	1597	81.12	9.99	48	73	83	90	95
Administrative Duties (%)	1597	11%						
Monthly Salary	1597	6389.81	2036.77	3042	4743	5962	7602	14464
% Female	1597	26%						
Highest Degree	1597							
Other	144	9.02%						
PhD	1350	84.53%						
Prof	103	6.45%						
Field	1597							
Arts	220	13.78%						
Other	1067	66.81%						
Prof	310	19.41%						
Rank	1597							
Assist	315	19.72%						
Assoc	437	27.36%						
Full	845	52.91%						

a. By comparing the statistics derived in problem 2 to those derived for this problem, what can you guess about the faculty who have been employed the longest at the university?

Ans: The above descriptive statistics are based on a single measurement for each individual, while the descriptive statistics in Problem 2 had multiple records for the faculty who had been at the university the longest. This means that the descriptive statistics for variables that do not change over time (e.g., sex, field, highest degree, years since degree, and start year) will be biased towards the faculty who have been at the university the longest. For instance, the fact that problem 2 shows a lower proportion of females than problem 3 suggests that fewer women than men have been at the university a long time. That also explains why problem 2 shows lower averages for starting year and year since receiving degree: There were up to 20 records for long term faculty, but only 1 record for faculty starting in 1995. There does not seem to be any particular trend in the field of employment, and there is a slight trend toward the longterm faculty having more professional degrees than are present in the

newer hires. The variables that change over time (salary, rank, and administrative duties) will tend to reflect a lower value in problem 2 than in problem 3. This can be anticipated by considering the following: When considering the faculty employed in 1995, we will tend to have more of the junior faculty and fewer of the senior faculty who were also employed in 1975. Thus the data from the earliest years are biased toward the lower ranks and salaries.

b. Based on the descriptive statistics you obtained in this problem, are there any outliers that you would worry about for any of the variables?

Ans: As in problem 2, the salary descriptive statistics are suggestive of some outliers.

4. It is also common to provide descriptive statistics to assess whether subjects in the groups of greatest interest (in this case, groups defined by sex) were similar with respect to other variables. Provide suitable descriptive statistics to address this question. Might there be confounding of the analysis of an association between sex and salary by other variables? Briefly explain.

	Males					Females				
Variable	N	Mean	Std. Dev.	Min	Max	N	Mean	Std. Dev.	Min	Max
Year of Degree	1188	74.37	9.64	48	96	409	81.11	8.70	54	95
Starting Year	1188	79.62	10.17	48	95	409	85.47	8.02	57	95
Admin Duties %	1188	12%				409	8%			
Monthly Salary	1188	6731.64	2089.76	3130.588	14464	409	5396.91	1481.22	3042	11036
Highest Degree	1188					409				
Other	88	7.41%				56	13.69%			
PhD	1016	85.52%				334	81.66%			
Prof	84	7.07%				19	4.65%			
Field	1188					409				
Arts	140	11.78%				80	19.56%			
Other	780	65.66%				287	70.17%			
Prof	268	22.56%				42	10.27%			
Rank	1188					409				
Assist	170	14.31%				145	35.45%			
Assoc	299	25.17%				138	33.74%			
Full	719	60.52%				126	30.81%			

Ans: The above descriptive statistics suggest that men tend to have received their degree earlier, tend to have been hired earlier, tend to have more administrative duties, are more likely to have a PhD or professional degree, are less likely to be in fine arts, and are much more likely to be full professors. As all of these variables tend to be associated with salary level, it is quite possible there is confounding. The problem, of course, will be deciding which of these variables might be in the causal pathway of interest. The following descriptive statistics are how I would examine an association between salary and field or highest degree. Note that I do want to be able to establish that field and highest degree are associated with salary independent of sex. Hence I consider descriptive statistics within each sex separately (as well as combined). Note that there is a clear tendency for faculty with professional degrees to be paid more than faculty with PhDs, and a clear tendency for faculty in fine arts to be paid less.

Descriptive Statistics for Monthly Salary

	Obs	Mean	Std. Dev.	Min	Max
By Degree					
Both sexes					
Other	144	5516.13	1478.54	3464	11840
PhD	1350	6399.69	2037.24	3042	14464
Prof	103	7481.75	2161.64	3934	13414
Males					
Other	88	5758.96	1587.44	3720	11840
PhD	1016	6731.57	2084.52	3131	14464
Prof	84	7751.52	2149.46	4348	13414
Females					
Other	56	5134.55	1206.26	3464	9420
PhD	334	5390.15	1486.82	3042	11036
Prof	19	6289.06	1825.71	3934	9693
By Field					
Both sexes					
Arts	220	5278.08	1265.54	3414	9974
Other	1067	6291.64	1993.81	3042	13998
Prof	310	7516.67	2095.36	3362	14464
Males					
Arts	140	5488.11	1280.22	3720	9974
Other	780	6641.72	2062.06	3131	13998
Prof	268	7642.97	2118.37	3362	14464
Females					
Arts	80	4910.54	1158.08	3414	9420
Other	287	5340.21	1411.61	3042	11036
Prof	42	6710.79	1759.70	4292	10263

5. Provide descriptive statistics of the data for starting salaries for those subjects who were hired as assistant professors in 1990. Compare the distribution of such starting salaries in that year for men and women. How would you use this data to decide whether men and women hired in 1990 were treated equally? What summary measures might you use for comparisons, and what are their relative merits?

Variable	Obs	Mean	Std. Dev.	Min	25th%ile	Median	75th%ile	Max
Both sexes	56	3876.69	667.04	2724	3318	3818	4220	5698
Males	26	4071.65	725.02	2825	3620	4128	4660	5698
Females	30	3707.72	571.91	2724	3263	3528	4058	5010

Ans: The above descriptive statistics address the starting salaries of those men and women who were hired in 1990 and are still at the university. There may well have been some faculty hired in that year who have since left the university. As it is quite possible that such faculty might have been paid

differently than those who stayed, I would not make much use of this data to answer the stated question. I would instead try to get appropriate data which was a random sample (or census) of all faculty hired in that year.