Biost 517: Applied Biostatistics I
Emerson, Fall 2012

Homework #1 Key
September 24, 2012

Written problems: To be handed in at the beginning of class on Wednesday, October 3, 2012 (See the end of this handout for the Data Analysis problem to be discussed in Discussion Section October 1-5.)

Homework papers will generally be returned during discussion section. Please indicate on your homework the discussion section you will likely attend next.

> *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.)*

Questions for Biost 514 and Biost 517:

The class web pages contains a description of a dataset regarding an observational study of cognitive function in the elderly (dsst.doc and dsst.txt). For this homework, we are only interested in the 7 variables measuring patient **age**, sex (**male**), **weight, height**, cognitive function (**dsst**), length of follow-up (**daysfu**), and vital status at the end of follow-up (**death**). Of particular interest for this homework is differences in measured cognitive function by sex, as well as any patterns in the covariates according to whether the cognitive function was missing or not.

General Comments:

The webpages contain a file containing annotated Stata code that was used to answer this question. In that document I provide some motivation for the variables I used to explore that time to death. In particular, I describe the generation of two variables that record

- **the number of days each subject was alive during the first 1201 days after study accrual, and**
- **an indicator of whether each subject died during those first 1201 days.**

Such data are "complete" (i.e., nonmissing) for every subject, as noted below.

Also included in the annotated Stata code is the generation of a variable that categorizes the DSST measurements in order to be able to explore comparisons between people missing DSST and people having available DSST measurements that are very low. In real life, when confronted with the missing DSST data, I did anticipate from the outset that such missing data might actually reflect subjects who were truly incapable of completing the test. Hence, from the outset I described the data using a dichotomization similar to the one I used here. However, in writing up this homework key, I pretended that it only occurred to me after seeing the comparison of subjects with complete or missing DSST scores. So I acknowledge that I performed some comparisons that are not shown, and then just show the data that I find most informative.

Also provided on the course webpages is an Excel spreadsheet that I used to format the data. For each table

- **The relevant table from the Stata format was "cut and pasted" into cell A40 of the corresponding worksheet.**
- **"Text to Columns" on the Excel Data menu was used to separate the copied tables into individual columns.**
- **Excel formatting code using TEXT( ) was used to create the entries for each row of the resulting table. My basic process was to**
    - **Create the first row of each table.**
    - **Copy that row into the next several rows to reflect all of the variables for the females.**
    - **I edited the number of significant digits to be displayed by changing the format argument supplied to the TEXT() commnd: "##0.0" provides one significant digit after the decimal point' "##0.000" provides three significant digits, etc.**
    - **I then could copy the whole block of rows to reflect the output for Males and Total, etc.**
- **I then copy the table into MS-Word, and format the table with bold face, centering of columns, etc.**
    - **Note that I delete the rows that are not of as much interest, instead just reporting the summary variables in the text.**

1. Where relevant, provide descriptive statistics for each of these variables in the entire sample, as well as within groups defined by sex. The descriptive statistics should provide information on the number of valid observations, the number of missing observations, the mean, the standard deviation, the minimum, 25^{th} percentile, median, 50^{th} percentile, 75^{th} percentile, and the maximum, where such statistics are of scientific interest.

    a. Note that the participants were followed for varying times. Because of this, the proportion of observed deaths are not immediately interpretable. What is the minimum follow-up among the patients still alive at the time of data extraction? What proportion of patients died within that time frame?

Answer: The data set includes information on 3,660 subjects (2,133 women and 1,527 men) ranging in age from 65 to 97 years old. Descriptive statistics are shown in Table 1 for each sex separately, as well as for the combined sample. Weight and height measurements have negligible missing data, but 118 subjects (68 women and 50 men) are missing data for the digit symbol substitution test (DSST). Because length of follow-up for survival varied across subjects, the data was recoded to reflect survival only through the first 1201 days following accrual to the study, because that was the minimal time of observation for subjects still alive at the time of data abstraction.

The age distribution is approximately the same for women and men, with an average age of 74.8 years (SD 5.10 years) and 75.5 years (SD 5.39 years) , respectively. As might be expected the men tend to be heavier and taller on average. Among subjects with DSST scores available, the scores for males tended to be slightly lower for males than for females: average DSST of 37.7 points (SD 13.3 points) versus 40.6 points (SD 13.7 points).

Mortality within the first 1,201 days following study accrual was 8.9% overall in the sample, but was 6.3% for women and 12.6% for men. Hence during those first 1201 days, men survived 1,138 days (SD 202 days) on average, while women averaged 1,167 days (SD 151 days) survival during that period.

Table 1: Descriptive statistics for selected variables

		N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
Females	**Age (yrs)**	2133 (0)	74.8 (5.10)	74 (71, 78)	(65, 97)
	Male	2133 (0)	0.000 (0)	0 (0, 0)	(0, 0)
	Weight (lbs)	2132 (1)	149 (30.0)	146 (128, 166)	(72, 320)
	Height (cm)	2132 (1)	159 (6.30)	159 (155, 163)	(124, 187)
	DSST	2065 (68)	40.6 (13.7)	40 (31, 50)	(0, 87)
	Days Alive During First 1201 Days	2133 (0)	1167 (151)	1201 (1201, 1201)	(66, 1201)
Males	**Age (yrs)**	1527 (0)	75.5 (5.39)	74 (71, 79)	(65, 97)
	Male	1527 (0)	1.000 (0)	1 (1, 1)	(1, 1)
	Weight (lbs)	1527 (0)	174 (27.1)	172 (156, 190)	(109, 300)
	Height (cm)	1527 (0)	173 (6.56)	173 (168, 177)	(153, 193)
	DSST	1477 (50)	37.7 (13.3)	38 (29, 47)	(0, 84)
	Days Alive During First 1201 Days	1527 (0)	1138 (202)	1201 (1201, 1201)	(11, 1201)
Total	**Age (yrs)**	3660 (0)	75.1 (5.23)	74 (71, 78)	(65, 97)
	Male	3660 (0)	0.417 (0)	0 (0, 1)	(0, 1)
	Weight (lbs)	3659 (1)	159 (31.3)	158 (137, 178)	(72, 320)
	Height (cm)	3659 (1)	165 (9.42)	164 (158, 172)	(124, 193)
	DSST	3542 (118)	39.4 (13.6)	40 (30, 49)	(0, 87)
	Days Alive During First 1201 Days	3660 (0)	1155 (174)	1201 (1201, 1201)	(11, 1201)

2. Comment on what the results of your analysis might say about the differences between those participants whose DSST measurements are available, and those whose DSST measurements are missing. Can you speculate on possible reasons for the trends you note?

Answer: The data were explored to examine any differences that might exist between those subjects who were or were not missing data for the DSST scores. In such comparisons, it was noted that compared to subjects with complete DSST data, subjects missing data for DSST tended to be older and to have a higher mortality during the first 1201 days following study accrual. Based on these results it was conjectured that subjects might be missing DSST scores because they were unable to perform the test. To further explore patterns of such missing data, descriptive statistics were computed within groups defined by having an observed DSST score of 10 or higher, having an observed DSST score less than 10, or having missing data for DSST. These descriptive statistics are presented in Table 2.

Among 3,479 subjects having an observed DSST of 10 or greater, 41.5% were male and the average age was 74.9 years (SD 5.03 years). Among 63 subjects having an observed DSST less than 10, 50.8% were male and the average age was 79.8 years (SD 6.66 years). The 118 subjects missing DSST data were 42.4% male and similar in age distribution to those subjects having low observed DSST scores with an average age of 79.9 years (SD 6.55 years).

Differences in mortality during the first 1201 days following study accrual were also evident across these groups: Subjects with observed DSST scores of 10 or greater had 7.9% mortality, subjects with observed DSST scores less than 10 had 25.4% mortality, and subject missing DSST scores had 29.7% mortality.

The observed similarity between subjects with missing DSST and those with low measurements for DSST suggest (but do not prove) that the predominant reason for missing DSST scores is an inability to perform the test.

Table 2: Descriptive statistics for selected variables within groups defined by DSST scores.

		N (msng)	Mean (SD)	Mdn (IQR)	(Min, Max)
DSST >= 10	**Age (yrs)**	3479 (0)	74.9 (5.03)	74 (71, 78)	(65, 97)
	Weight (lbs)	3478 (1)	160 (31.3)	158 (137, 178)	(72, 320)
	Height (cm)	3478 (1)	165 (9.43)	164 (158, 172)	(124, 193)
	DSST	3479 (0)	40.0 (12.9)	40 (31, 49)	(10, 87)
	Days Alive During First 1201 Days	3479 (0)	1160 (164)	1201 (1201, 1201)	(11, 1201)
DSST <10	**Age (yrs)**	63 (0)	79.8 (6.66)	80 (75, 85)	(66, 97)
	Weight (lbs)	63 (0)	158 (33.4)	162 (130, 184)	(97, 255)
	Height (cm)	63 (0)	165 (9.21)	164 (158, 172)	(148, 184)
	DSST	63 (0)	4.6 (3.2)	5 (2, 7)	(0, 9)
	Days Alive During First 1201 Days	63 (0)	1114 (214)	1201 (1201, 1201)	(340, 1201)
DSST missing	**Age (yrs)**	118 (0)	79.9 (6.55)	80 (74, 85)	(68, 97)
	Weight (lbs)	118 (0)	151 (31.6)	148 (128, 166)	(86, 284)
	Height (cm)	118 (0)	163 (9.21)	161 (157, 170)	(142, 190)
	DSST	0 (118)	--	--	--
	Days Alive During First 1201 Days	118 (0)	1024 (323)	1201 (1017, 1201)	(54, 1201)
Total	**Age (yrs)**	3660 (0)	75.1 (5.23)	74 (71, 78)	(65, 97)
	Weight (lbs)	3659 (1)	159 (31.3)	158 (137, 178)	(72, 320)
	Height (cm)	3659 (1)	165 (9.42)	164 (158, 172)	(124, 193)
	DSST	3542 (118)	39.4 (13.6)	40 (30, 49)	(0, 87)
	Days Alive During First 1201 Days	3660 (0)	1155 (174)	1201 (1201, 1201)	(11, 1201)