

Biost 517: Applied Biostatistics I

Emerson, Fall 2011

Homework #5 Key

October 29, 2011

Written problems: To be handed in at the beginning of class on Wednesday, November 2, 2011.

*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 following problems make use of a dataset exploring the association between cerebral changes seen on head MRI and all cause mortality. The documentation file mri.doc and the data file mri.txt can be found on the class web pages.

1. Consider the censoring distribution for this dataset.
 - a. Provide suitable statistics for the distribution of times to censoring for observations of death.

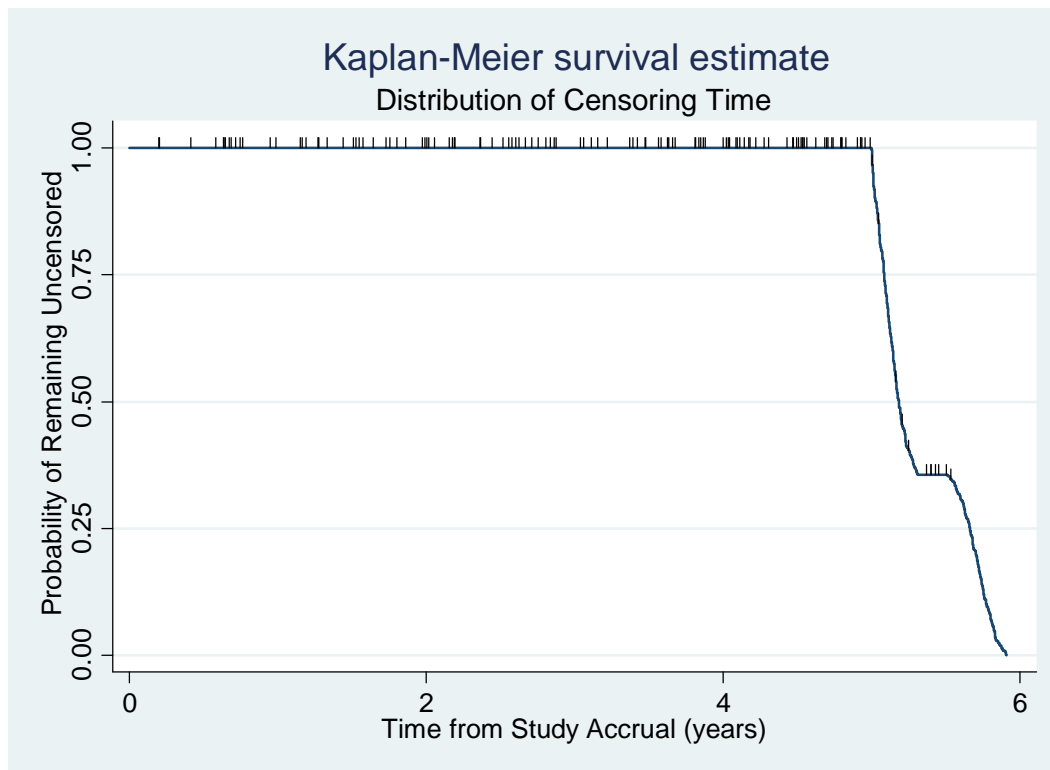


Table 1: Descriptive statistics for the distribution of time to censoring in months. Statistics provided include the number of observations (N), the number of observed events (Ev), the mean (possibly restricted to a specified time (Restr), the median (Mdn), the 25th and 75th percentiles (interquartile range or IQR), the minimum (Min)

and maximum (Max) (possibly censored as denoted with an asterisk (*)), and the probability of being uncensored at 6, 12, and 18 months.

	N (Ev)	Mean (Restr)	Mdn (IQR)	(Min, Max)	(5, 5.5, 6 yr Uncens)
Censored	735(602)	5.33 (NA)	5.19 (5.09, 5.66)	(0.186*, 5.91)	(1.000, 0.355, 0.000)

Ans: From the above figure and Table 1, it can be seen that the observed censoring times occur between 5 and 6 years from the time of study accrual, with a mean of 5.33 years and median of 5.19 years. The Kaplan-Meier estimate for the probability of remaining uncensored by 5 years is 100%. (In fact, the only censoring in this study is from administrative censoring due to the patient being still alive when the dataset was defined. Hence, the true censoring distribution is known absolutely from the protocol and the time of study accrual. However, there is nothing in this dataset that would have told you that. You would have needed to have the protocol and the accrual times.)

- b. Suppose we want to divide individual patients into groups who die within 5 years and those who do not. On the basis of your answer to part a, will we be able to do so?

Ans: As the earliest observed censoring was at 5.002 years, we do know for each subject whether they survived at least 5 years or not. Hence, we can easily divide this dataset into groups based on 5 year survival.

- 2. Suppose we are interested in using the scores on the digit symbol substitution test (DSST) to predict whether a patient will be dead within five years after study accrual.

Ans: Because the only sample size that was constrained in this study was the total sample size (it was a one sample cohort study, so it can be viewed as cross-sectional for our purposes), we can estimate the prevalence of “disease” (death within 5 years), the prevalence of a “positive test” (a DSST less than 35), the sensitivity, the specificity, the predictive value of a positive, and the predictive value of a negative directly from the cross tabulation of variables indicating death within 5 years and DSST lower than 35. The following table provides that cross tabulation.

Table 2: Crosstabulation of subjects in the data set according to observed death within 5 years and a DSST score less than or greater than or equal to 35. Subjects with missing DSST scores are omitted from the analysis. Also included are percentages computed across rows (r) and across columns (c).

	DSST ≥ 35	DSST < 35	Total
Alive for 5y	450 (r 74.0%, c 89.1%)	158 (r 26.0%, c 72.5%)	608 (r 100%, c 84.1%)
Dead in 5y	55 (r 47.8%, c 10.9%)	60 (r 52.2%, c 27.5%)	115 (r 100%, c 15.9%)
Total	505 (r 69.8%, c 100%)	218 (r 30.2%, c 100%)	723 (r 100%, c 100%)

- a. In our sample, what is the prevalence of death within 5 years?

Ans: From Table 2 we find that 15.9% of subjects died within 5 years. (In the above table, 12 patients were missing DSST scores, and hence they are excluded from this analysis. It should be noted that 6 of those 12 patients died. This 50% death rate in this subgroup is a little higher than would be expected if the missing data were missing completely at random, but such an observation is clearly not impossible. In any case, it is probably a better approach to calculate the prevalence using all available data on death within 5 years, but then there will be somewhat

of a very slight contradiction between the prevalence used for calculating the PPV and NPV and the prevalence calculated for all cases. There is clearly no easy solution. Anything we do is making some sort of assumption. I chose to use the same sample to calculate all of parts a-d, but there was no requirement that you do.)

- b. In our sample, what is the prevalence of a DSST less than 35? (For parts b, c, and d, you may just omit cases having a missing value for DSST.)

Ans: From Table 2 we find that 30.2% of subjects have a DSST less than 35.

- c. Suppose we consider a DSST less than 35 to be a “positive” test result. What are the sensitivity and specificity of such a diagnostic criterion? Briefly explain how these were calculated.

Ans: Sensitivity in this setting would be the percentage of subjects having a DSST lower than 35 among those who died within 5 years. From Table 2 we find a sensitivity of $60 / 115 = 52.2\%$.

Specificity in this setting would be the percentage of subjects having a DSST greater than or equal to 35 among those who survived at least 5 years. From Table 2 we find a specificity of $450 / 608 = 74.0\%$.

- d. If the sample accurately reflects the patient population of interest, what are the positive and negative predictive values of such a diagnostic criterion? Briefly explain how these were calculated.

Ans: Positive predictive value in this setting would be the percentage of subjects observed to die with 5 years among those having a DSST lower than 35. From Table 2 we find a positive predictive value of $60 / 218 = 27.5\%$.

Negative predictive value in this setting would be the percentage of subjects observed to survive at least 5 years among those having a DSST greater than or equal to 35. From Table 2 we find a negative predictive value of $450 / 505 = 89.1\%$.

- e. Now suppose that subjects who are missing scores for the DSST just refused to take the test because they found it too taxing. In such a situation we might consider a missing not at random (MNAR) model in which we “impute” their scores to be 0. Repeat parts b, c, and d with this imputed data.

Table 3: Crosstabulation of subjects in the data set according to observed death within 5 years and a DSST score less than or greater than or equal to 35. Subjects with missing DSST scores are imputed in a MNAR model to have a DSST of 0. Also included are percentages computed across rows (r) and across columns (c).

	DSST \geq 35	DSST < 35	Total
Alive for 5y	450 (r 73.3%, c 89.1%)	164 (r 26.7%, c 71.3%)	614 (r 100%, c 83.5%)
Dead in 5y	55 (r 45.4%, c 10.9%)	66 (r 54.6%, c 28.7%)	121 (r 100%, c 16.5%)
Total	505 (r 68.7%, c 100%)	230 (r 31.3%, c 100%)	735 (r 100%, c 100%)

Ans: From Table 3 we find that 16.5% of subjects died within 5 years, and 31.3% of subjects had a DSST that was less than 35 or was missing.

From Table 3 we find a sensitivity of $66 / 121 = 54.6\%$ and a specificity of $450 / 614 = 73.3\%$.

From Table 3 we find a positive predictive value of $66 / 218 = 28.7\%$ and a negative predictive value of $450 / 505 = 89.1\%$.

(Note: In an analysis of a larger set of data from this study, I did find that those subjects whose DSST scores were missing tended to have survival approximately the same as subjects whose DSST scores were measured as 0. This does not prove that the above MNAR imputation model was correct. Hence, in the face of missing data, an analysis as was performed in this part is best regarded as a “sensitivity analysis” investigating the impact of possible missing data mechanisms. When you have missing data, there is no analysis that you can do that is not making some presumption about the missing data. The analyses in parts b-d presumed a missing completely at random (MCAR) mechanism, while this analysis presumed MNAR.)

- f. Suppose that the sample that we obtained undersampled patients who would actually die. If the true prevalence of death within five years in the target population were 25%, what would be the positive and negative predictive values of the diagnostic criterion based on a DSST less than 35 when we impute the missing data as 0? Briefly explain how these were calculated.

Ans: We typically presume that the sensitivity and specificity would generalize to a population with a different prevalence. Hence we use the sensitivity and specificity calculated in part e along with the hypothesized prevalence and Bayes Rule to find positive and negative predictive values as given below.

$$\begin{aligned}
 PPV = \Pr(D | +) &= \frac{\Pr(+ | D)\Pr(D)}{\Pr(+ | D)\Pr(D) + \Pr(+ | H)\Pr(H)} = \frac{Sens \times Prev}{Sens \times Prev + (1 - Spec) \times (1 - Prev)} \\
 &= \frac{0.546 \times 0.25}{0.546 \times 0.25 + (1 - 0.733) \times (1 - 0.25)} = 40.5\%
 \end{aligned}$$

$$\begin{aligned}
 NPV = \Pr(H | -) &= \frac{\Pr(- | H)\Pr(H)}{\Pr(- | H)\Pr(H) + \Pr(- | D)\Pr(D)} = \frac{Spec \times (1 - Prev)}{Spec \times (1 - Prev) + (1 - Sens) \times Prev} \\
 &= \frac{0.733 \times (1 - 0.25)}{0.733 \times (1 - 0.25) + (1 - 0.546) \times 0.25} = 82.9\%
 \end{aligned}$$