January 13, 2014

Biost 518

Homework #1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | #1(5) | #2(10) | #3(10) | #4(10) | #5(10) | #6(10) | #7(10) | #8(10) | Total(75) |
| point | 1 | 6.8 | 9 | 0 | 6 | 5 | 8 | 0 | 35.8 |

1.1/5

In order to provide descriptive statistics, we first generated a new variable obstime\_month representing

the observed time in month (we divided the obstime variable by 30.4). Then we set our data to survival

time by using the “stset” command and produced the Kaplan Meier survival estimates at 1, 2,3,4 and 5

years respectively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Time in months | Total at the beginning | Total Failed | Survivor Function | Std. Error | [95% Confidence Interval] |
| 12 months  | 722 | 14 | 0.98 | 0.005 | 0.9680 | 0.9887 |
| 24 months  | 702 | 20 | 0.95 | 0.008 | 0.9359 | 0.9667 |
| 36 months | 678 | 24 | 0.92 | 0.010 | 0.8991 | 0.9384 |
| 48 months | 654 | 24 | 0.89 | 0.012 | 0.8634 | 0.9091 |
| 60 months | 615 | 39 | 0.84 | 0.014 | 0.8065 | 0.8603 |

The above table provides descriptive statistics of the survival distribution and shows that 84% of the

people have a 5 year survival probabilities meaning that 15%of them were censored by 5

years. These are exactly the same as when we tabulate the dichotomized obstime variable (table below),

therefore it makes senses to dichotomize the time to death according to death within or after 5 years of

study enrollment.

|  |  |  |
| --- | --- | --- |
| Observed time  | N | Percentage |
| Less than 5 years | 121 | 16.46 |
| 5 years or more | 614 | 83.54 |
| Total | 735 | 100.00 |

The statement is NOT clear. Please state that why you think your Kaplan Meier result match the dichotomized one. Please check the definition of censoring again. In this problem, within 5 years, the information all holds(when a subject has the status 1, he/she is dead. When a subject has a status 0, he/she is alive.). No censoring within 5 years. So that’s why we can dichotomize rather than using Kaplan Meier method.

You will get 1 point for your effort on this. You miss 4 points because you are not on the right direction of solving this problem.

2.6.8/10

The table below shows the characteristics of the study sample stratified by low density lipropotein (LDL)

serum levels . Examining the two groups data reveals that there is low variability between the two groups:

the proportions, mean, standard deviations and ranges for the two groups are very similar.

|  |  |
| --- | --- |
|  | Low density lipoprotein (LDL) levels in serum |
|  | Not high | High |
| Observed time (in months) | n (%) | Mean | SD | Min | Max | n (%) | Mean | SD | Min | Max |
|  < 5 years | 105 (16.99%) |  |  |  |  | 16 (13.68%) |  |  |  |  |
|  ≥ 5 years | 513 (83.01%) |  |  |  |  | 101 (86.32%) |  |  |  |  |
| Age (years) |  | 74.51 | 5.4 | 65 | 99 |  | 74.84 | 5.78 | 65 | 94 |
| Weight (in pounds) |  | 159.36 | 30.78 | 86 | 264 |  | 163.09 | 30.45 | 74 | 257 |
| Smoking history (in pack-years) |  | 19.88 | 27.62 | 0 | 240 |  | 18.09 | 24.26 | 0 | 102 |
| Prior coronary heart disease |  |  |  |  |  |  |  |  |  |  |
| No  | 488 (78.96%) |  |  |  |  | 92 (78.63%) |  |  |  |  |
| Diagnosis of angina | 54 (8.74%) |  |  |  |  | 10 (8.55%) |  |  |  |  |
| Diagnosis of myocardial infarction | 76 (12.30%) |  |  |  |  | 91 (12.38%) |  |  |  |  |
| Prior congestive heart failure |  |  |  |  |  |  |  |  |  |  |
| No  | 581 (94.01%) |  |  |  |  | 113 (96.58%) |  |  |  |  |
| Yes | 37 (5.99%) |  |  |  |  | 4 (3.42%) |  |  |  |  |
| Prior cerebrovascular event  |  |  |  |  |  |  |  |  |  |  |
| No | 541 (87.54%) |  |  |  |  | 95 (81.2%) |  |  |  |  |
| Diagnosis of a transient ischemic attack | 18 (2.91%) |  |  |  |  | 6 (5.13%) |  |  |  |  |
| Diagnosis of a stroke | 59 (9.55%) |  |  |  |  | 16 (13.68%) |  |  |  |  |

You may want to the following revise:

1. include the missing # of data. As is mentioned in the problem, some data(marked as NA) is missing. And sample size is also an important information we need to know(precision? The proper analyzing method? And so on )

2. for binary variable, please include the mean value. It is the proportion of the “1” event. You may need this to compare over 2 groups.

3. please consider why you need these descriptive statistics. What you want to get from the statistics should be in your “result” part. You did not include the result in your answer, please check the key.

4. For the “method” part, you may want to clarify your grouping variable and outcome variable. And mention what statistics you assign to them.

You get 3.8/4 for the table layout.(0.2 point missing because the categorical variable part is a little bit messy.) 2.9/3 for the choice of descriptive statistics.(you got 0.1 missing because you did not include the count/missing count, mean for binary variable).0.1/3for the method and result finding.(please note that it’s hard to define “similar”. You might need to consider more in detail what does the data mean. In descriptive statistics, we need to pay more attention on the fact. You might want to state the values you get. It is not convincing to give your opinion(similar) without any evidence.)

3. 9/10

We performed a two sample t-test with unequal variance to assess whether having low LDL levels in

serum negatively affects the 5 year survival probability among elderly adults. The null hypothesis used

for the analysis is that the mean LDL levels in serum is equal for adults with and without a 5 year survival

probabilities. We are testing a one –sided alternative hypothesis based on the scientific question whether

low LDL levels in serum are associated with death from any cause. Based on our results, the difference in

LDL levels in serum among people with or without 5 year survival probabilities is 8.50 mg/dL. We are

95% confident???? that people with survival probability below 5 years have between 1.44 mg/dL and 15.56

mg/dL less LDL in serum than people with survival probability above 5 years. Based on the one sided p-

value of 0.0093, we reject the null hypothesis and conclude that low LDL levels in serum are statistically

associated with survival probabilities below 5 years.

1. I’m not sure if we can interpret as “95% confident….” But you should keep in mind that it’s NOT a probability that the true value lies between the intervals.
2. the problem is more general(association), you consider it in a more specific way(negatively). But I think it’s OK.

You get 9 point for this question. The missing 1 point is the scientific wording you might need to work on.

4.0/10

5.6/10

We are interested at testing the association between serum LDL levels and 5 years all cause mortality

among an elderly population. We conducted a chi squared test to assess whether having low LDL levels

in serum negatively affects the 5 year survival probability among elderly adults. Our null hypothesis

states that the probability of surviving 5 years or more is the same among those with high and low levels

of serum LDL. The alternative is that the probability of surviving 5 years or more is less among people

with low levels of serum LDL.

The p-value for the chi-squared test is 0.375 and greater than 0.05, we can’t reject the null hypothesis that

those with the probability of surviving 5 years or more is the same among those with high and low levels

of serum LDL at the 0.05 level.

You get 6 point for this question because you choose the right way to test and finished the test . You get 4 point missing because (1)you just give a P value in this problem. Please include more information in your result part.(for example, the value in your 2\*2 table.)(2) scientific wording.

6. 5/10

We calculated the odd ratio using a two-by-two table and then computed the confidence interval. The

odds ratio of surviving less than 5 years is 29% higher (CI: 0.74-2.26) for people with low serum LDL as

compared to the group with higher serum LDL levels. Unfortunately, the precision is not adequate to

demonstrate that the difference would be unlikely in the absence of true association (p=0.3753).

 You got 5 point for this problem. 5 points missing (1) fail to clarify the method you use(2) missing odds ratio information(3)scientific wording.

7.8/10



A log rank test was to assess the evidence that low serum LDL levels affects time to death or end of

study. Based on the two-sided p-value of 0.2664, we can’t reject the null hypothesis that there is no

difference in the probability of surviving 5 years or more between the low and high serum LDL levels.

|  |  |  |
| --- | --- | --- |
| LDL serum levels | Events observed | Events expected |
| < 160 mg/dL | 116 | 111.26 |
|  160 mg/dL | 17 | 21.74 |
| Total | 133 | 133 |

You get 8 point for this because you choose the right method and finish the test. The missing point is again some information missing and improper scientific wording. Please check the key as a template.

8.0/10

.