BIOSTATS 518

Due Monday, January 12th, 2015

1. The observations of time to death in this data are subject to (right) censoring. Nevertheless, problems 2 – 6 ask you to dichotomize the time to death according to death within 4 years of study enrolment or death after 4 years. Why is this valid? Provide descriptive statistics that support your answer.

The minimum amount of time in the study for those who did not die (death = 0) was 1480 days, which is more than 4 years. Thus, anyone that died before 4 years had observed deaths, and there was no censoring that occurred before 4 years. By dichotomizing using a cut off point of 4 years (or 1460 days, which is less than 1480 days), we can say with certainty that we know who died before that time cut off and who survived at least until that cut off.

1. Provide a suitable descriptive statistical analysis for selected variables in this dataset as might be presented in Table 1 of a manuscript exploring the association between serum CRP and 4 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease.

**Methods:** Binary indicator variables were used for previous smoking history, and previous cardiovascular disease. Descriptive statistics were calculated by subgroupings of CRP level with levels of <1 mg/L, 1-3 mg/L, >3 mg/L, and the entire sample. Continuous variables (age, BMI, and cholesterol) were described using mean, standard deviation, minimum and maximum, and binary variables (sex, smoking history, and prior cardiovascular disease) were described using percentages.

**Results:** There were 68 missing values in for CRP level in the data set, which were excluded from this analysis. Analysis was done on 4933 (those of the 5000 subjects without missing CRP values). There was some missing data for others variables of interest, however, it was assumed these values were missing at random and so these values were not omitted from analysis. There were 428 subjects with serum CRP levels below 1 mg/L, 3330 subjects with serum CRP levels between 1 and 3 mg/L, and 1175 subjects with serum CRP levels above 3 mg/L. Across all serum CRP level intervals, subjects were less likely to be male. Mean age did not vary significantly between serum CRP level intervals. Mean BMI tended to increase as CRP level increased. Those subjects with higher serum CRP levels were more likely to be smokers than those with lower serum CRP levels, and a higher proportion of those subjects with higher serum CRP levels had prior occurrence of cardiovascular disease. Those with higher serum CRP levels had a higher proportion of death within 4 years of study enrollment.

|  |  |
| --- | --- |
|  | **Serum C Reactive Protein (CRP) Level** |
|  | Below 1 mg/L(n = 428) | 1-3 mg/L(n = 3330) | Above 3 mg/L(n = 1175) | Any Level(n = 4933) |
| **Male (%)** | 45.6% | 43.3% | 37.0% | 41.9% |
| **Age (yrs) (mean (sd, min-max))** | 73.5 (5.8, 65-94) | 72.7 (5.5, 65-100) | 72.9 (5.7, 65-93) | 72.8 (65-100) |
| **BMI (mean (sd, min-max))** | 23.8 (3.6, 15.6-38.6) | 26.4 (4.3, 14.7-53.2) | 28.4 (5.5, 15.3-58.8) | 26.7 (4.7, 14.7-58.8) |
| **Smoker (%)** | 9.6% | 11.0% | 15.9% | 12.1% |
| **Cholesterol (mean (sd, min-max))** | 206.0 (40.5, 109-407) | 212.8 (38.6, 73-363) | 210.6 (40.6, 97-430) | 211.7 (39.3, 73-430) |
| **Prior Cardiovascular Disease (%)** | 18.2% | 21.5% | 28.7% | 23.0% |
| **Death within 4 years** | 4.9% | 8.4% | 15.6% | 9.8% |

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing mean CRP values across groups defined by vital status at 4 years.

**Methods:** Mean serum C reactive protein (CRP) levels were compared between those subjects who died before 4 years in the study and those who survived at least 4 years in the study. A t test allowing for unequal variances was used to determine differences in mean and 95% confidence intervals.

**Results:** Mean serum CRP level was 5.38 mg/L for the 484 subjects that did not survive beyond 4 years of study enrollment, whereas mean CRP level was 3.42 mg/L for the 4449 subjects that survived past 4 years, with a difference in mean CRP level between the two groups of 1.95 mg/L. The 95% confidence interval indicates that the observed difference would not be unusual if the true difference in mean CRP levels between those subjects that died before 4 years and those that survived past 4 years in study enrollment fell between 2.70 mg/L and 1.21 mg/L. Using an alpha level of 0.05, we can reject the null hypothesis that there is no difference in mean serum CRP level between subjects who died before 4 years of study enrollment and those that survived past 4 years (P < 0.001).

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing geometric mean CRP values across groups defined by vital status at 4 years. (Note that there are some measurements of CRP that are reported as zeroes. Make clear how you handle these measurements.)

**Methods:** Geometric means of C reactive protein were compared between those subjects who died before 4 years in the study and those who survived at least 4 years in the study. To account for CRP levels of zero, values of zero were replaced with half of the minimum non-zero CRP level, which was 1 (so zeros were replaced with 0.5). A t test allowing for unequal variances was performed on the log of CRP values. Geometric means, and differences in geometric means, as well as a 95% confidence interval, was determined after exponentiating the values determined by the t test.

**Results:** Geometric mean serum CRP level was 2.03 mg/L for the 484 subjects that did not survive beyond 4 years of study enrollment, whereas geometric mean CRP level was 0.97 mg/L for the subjects that survived past 4 years, with a 32% higher geometric mean CRP level for subjects that did not survive beyond 4 years (the difference in geometric mean CRP level between the two groups of 0.68 mg/L). The 95% confidence interval indicates that the observed difference would not be unusual if the true population geometric mean CRP levels between those subjects that died before 4 years and those that survived past 4 years in study enrollment fell between 38% and 25% higher CRP level. Using an alpha level of 0.05, we can reject the null hypothesis that there is no difference in geometric mean serum CRP level between subjects who died before 4 years of study enrollment and those that survived past 4 years (P < 0.001).

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods:** the proportion of subjects who did not survive beyond 4 years of study enrollment was compared between those who had high serum CRP levels to those who did not. Differences in probability of not surviving beyond 4 years of study enrollment were tested using Pearson’s chi squared test for independence, and 95% confidence intervals were calculated using Wald statistics.

**Results:** Of subjects without high serum CRP levels (CRP ≤ 3 mg/L, n = 3758), 8.0% were observed to die within 4 years of study enrollment, whereas 15.6% of subjects with high serum CRP levels (CRP > 3 mg/L, n = 1175) were observed to die within 4 years of study enrollment. The 95% confidence interval indicates that the observed difference in probability of death of 7.6% between the two groups would not be unusual if the true probability of death within fell between 5.3% higher and 9.8% higher for those with high serum CRP levels compared to those without high serum CRP levels. Using a chi-squared test with an alpha level of 0.05, this observation is statistically significant and thus we can reject the null hypothesis that survival probabilities are not associated with serum CRP levels (p < 0.001).

1. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the odds of death within 4 years across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods:** Odds of death within 4 years of study enrollment were compared between the group of subjects were high serum CRP level (>3 mg/L) and those without high serum CRP level (≤ 3 mg/L). Differences in odds of not surviving beyond 4 years of study enrollment were tested using Pearson’s chi squared test for independence, and 95% confidence intervals were calculated using Wald statistics.

**Results:** Odds of death within 4 years of study enrollment for those without high serum CRP level was 0.09, whereas odds of death within 4 years of study enrollment for those with high serum CRP level was 0.18. Based on the 95% confidence interval, the observed odds ratio of 2.12 comparing the high serum CRP group to the lower serum CRP group would not be considered unusual if the true odds ratio fell between 1.74 and 2.57. Using a chi-squared test with an alpha level of 0.05, this observation is statistically significant and thus we can reject the null hypothesis that odds of death within 4 years of study enrollment are not associated with serum CRP levels (p < 0.001).

1. Perform a statistical analysis evaluating an association between serum CRP and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum CRP (“high” = CRP > 3 mg/L).

**Methods:** Kaplan-Meier survival estimates were calculated for subjects grouped by serum CRP level, which high serum CRP level defined as >3 mg/L and lower serum CRP level defined as ≤3 mg/L. Difference in survival probabilities between serum CRP level groups were tested using the logrank test, with a 95% confidence interval computed using the Cox proportional hazards method.

**Results:** As is shown in the Kaplan-Meier survival graph below, those subjects with lower serum CRP level (≤3 mg/L) have higher survival probabilities at all points in time in the study. The instantaneous risk of death is 69% higher for those with high serum CRP levels (>3 mg/L). Based on a 95% confidence interval, the observed hazard ratio of 1.69 would not be unusual if the true hazard ratio fell between 1.49 and 1.92. Using a logrank test with a p-value < 0.001 indicates that these results are statistically significant and we can reject the null hypothesis that there is no difference in instantaneous risk of death between serum CRP level groups.

|  |  |
| --- | --- |
|  | **Kaplan-Meier Survival Probabilities** |
|  | Serum CRP Level ≤ 3 mg/L | Serum CRP Level > 3 mg/L |
| **1 year** | 0.988 | 0.967 |
| **2 years** | 0.971 | 0.926 |
| **3 years** | 0.984 | 0.881 |
| **4 years** | 0.920 | 0.844 |
| **5 years** | 0.884 | 0.800 |

1. Supposing I had not been so redundant (in a scientifically inappropriate manner) and so prescriptive about methods of detecting an association, what analysis would you have preferred *a priori* in order to answer the question about an association between mortality and serum CRP? Why?

*A priori*, and before seeing results, for this set of data given this study design, a t test comparing geometric mean CRP levels would have been most appropriate. Considering serum CRP level as an exposure of interest, and death before 4 years of study enrollment as an outcome of interest makes the most sense because serum CRP levels are set prior to death (exposure precedes the outcome in time). Looking at geometric mean CRP levels in each group makes more sense than dichotomizing the data, given that CRP levels are a continuous variable, and reporting geometric means is therefore more precise. Given the range of CRP levels in the data set (0-48), with cut offs of <1, 1-3, and >3 mg/L, we can see that the scale of biological/scientific interest is not proportional to the range of data, and thus looking at the data on a multiplicative scale instead of looking at absolute differences makes more sense in this case. Thus, using geometric means to compare instead of means is probably more precise. Comparing geometric means between these groups is the most simple and valid way to convey the association between mortality and serum CRP.