**Homework #1**

1. In order to determine why it is valid to dichotomize the time to death according to death within 5 years of study enrollment, we cross-tabulated the binary variables which indicated whether or not a patient survived greater than 5 years after the start of enrollment and whether or not a patient was observed to die during the study. If the patient was not observed to die during the study because the study ended before they died, then that patient’s time to death was censored.

|  |  |  |
| --- | --- | --- |
|  | Observed to Die During Study (Uncensored Time to Death) | Not Observed to Die During Study (Censored Time to Death) |
| Survived Less Than 5 Years After Start of Enrollment | 121 | 0 |
| Survived Greater Than 5 Years After Start of Enrollment | 12 | 602 |

It is valid to dichotomize the time to death according to death within 5 years of study enrollment because all right censoring of the observations of time to death occur for patients who have already been enrolled for at least 5 years. This is clear from the above table, which shows that for each of the 121 patients who survived less than 5 years after the start of enrollment, death was observed during the study, so there was no censoring of the time to death in this population. By contrast, of the 614 patients who survived greater than 5 years after the start of enrollment, only 12 patients were observed to die during the study, while the remaining 602 patients were not observed to die during the study. So for all 602 of these patients, their time to death is censored, but we can say with certainty that it is greater than 5 years.

2. **Table 1: Descriptive Statistics Dichotomized by LDL Level**

|  |  |  |
| --- | --- | --- |
|  | **Low LDL (LDL < 160 mg/dL)****n=618** | **High LDL (LDL ≥ 160 mg/dL)****n=117** |
| **Variable****(units, number missing)** | **n (%)** | **Mean** | **SD** | **Min** | **Max** | **Median** | **n (%)** | **Mean** | **SD** | **Min** | **Max** | **Median** |
| **Age (in years)** |  | 74.51 | 5.39 | 65 | 99 | 73 |  | 74.84 | 5.78 | 65 | 94 | 74 |
| **Weight (in lbs)** |  | 159.4 | 30.78 | 86 | 264 | 158 |  | 163.1 | 30.45 | 74 | 257 | 158 |
| **Pack Years (in years, 1 missing)** |  | 19.88 | 27.62 | 0 | 240 | 7 |  | 18.09 | 24.26 | 0 | 102 | 3.75 |
| **Years Since Quitting Smoking (in years)** |  | 9.654 | 14.17 | 0 | 56 | 0 |  | 9.701 | 13.81 | 0 | 56 | 0 |
| **Gender (Male=1)** |  | 0.510 |  | 0 | 1 |  |  | 0.436 |  | 0 | 1 |  |
| **History of CHF (Yes=1)** |  | 0.060 |  | 0 | 1 |  |  | 0.034 |  | 0 | 1 |  |
| **History of CHD** |  |  |  |  |  |  |  |  |  |  |  |  |
| **No History** | 488(79%) |  |  |  |  |  | 92(79%) |  |  |  |  |  |
| **Diagnosis of Angina** | 54(9%) |  |  |  |  |  | 10(9%) |  |  |  |  |  |
| **Diagnosis of MI** | 76(12%) |  |  |  |  |  | 15(13%) |  |  |  |  |  |
| **History of Stroke** |  |  |  |  |  |  |  |  |  |  |  |  |
| **No History** | 541(88%) |  |  |  |  |  | 95(81%) |  |  |  |  |  |
| **Diagnosis of TIA** | 18(3%) |  |  |  |  |  | 6(5%) |  |  |  |  |  |
| **Diagnosis of Stroke** | 59(10%) |  |  |  |  |  | 16(14%) |  |  |  |  |  |
| **Survived Greater Than 5 Years (Yes=1)** |  | 0.830 |  | 0 | 1 |  |  | 0.863 |  | 0 | 1 |  |

We chose to dichotomize Table 1 by high LDL (henceforth defined as LDL ≥ 160 mg/dL) and low LDL (henceforth defined as LDL < 160 mg/dL) because the most clinically relevant manner to approach this data is to view 5-year all-cause mortality as our response variable and to view the presence of high LDL levels as our predictor of interest for the response variable.

3. To evaluate whether or not there is an association between LDL and 5-year mortality, we performed a two-sample t-test, without assuming equal variance, for the equality of mean LDL levels between patients who survived for less than 5 years after enrollment and those who survived greater than 5 years after enrollment. This test compared mean LDL levels in the patient groups defined previously by taking their difference.

After performing the t-test, we found that for the 119 patients who survived for less than 5 years after enrollment, the mean LDL level was 118.7 mg/dL, and for the 606 patients who survived greater than 5 years after enrollment, the mean LDL level was 127.2 mg/dL, giving an average difference in mean LDL levels of 8.5 mg/dL higher for the group which survived greater than 5 years after enrollment. A 95% confidence interval for this difference is bounded below by 1.4 mg/dL higher for the group which survived greater than 5 years after enrollment and is bounded above by 15.6 mg/dL higher for the group which survived greater than 5 years after enrollment. This means with 95% confidence, it would not be unusual if the true mean LDL levels for the group which survived greater than 5 years after enrollment were between 1.4 mg/dL higher or 15.6 mg/dL higher than the mean LDL levels for the group which survived less than 5 years after enrollment. This t-test had a statistically significant two-sided p-value of 0.0186, which means that we can with high confidence state that this difference indicates an association between LDL and 5-year mortality.

4. To evaluate whether or not there is an association between LDL levels and 5-year mortality, we chose to compare the geometric means of LDL levels between patients who survived for less than 5 years after enrollment and those who survived greater than 5 years after enrollment. To compare the geometric mean LDL levels, we took the ratio of geometric mean LDL levels between groups. We did this by taking the natural logarithm of all LDL levels, and then performing a two-sample t-test (without assuming equal variance) for the equality of mean log(LDL) levels between the two groups of interest defined previously, which produced a mean difference in log(LDL) levels and a 95% confidence interval for this difference. To derive the ratio of the geometric mean LDL levels between these groups, we raised the mean log(LDL) difference and both endpoints of the 95% confidence interval as powers of the natural base e. This ratio will be the geometric mean LDL level for the group surviving less than 5 years after enrollment divided by the geometric mean LDL level for the group surviving greater than 5 years after enrollment.

The two-sample t-test as described above produced a mean difference in log(LDL) levels of -0.092, where the mean log(LDL) level for the group surviving less than 5 years after enrollment was equal to 4.719 (corresponding to a geometric mean LDL level of 112.1) and the mean log(LDL) level for the group surviving greater than 5 years after enrollment was equal to 4.811 (corresponding to a geometric mean LDL level of 122.9). A 95% confidence interval for this difference is (-0.164, -0.020). After deriving the ratio of the geometric mean LDL levels between these groups as described above, we obtained a point estimate of the ratio of geometric means of 0.912, and a 95% confidence interval for this ratio of (0.848, 0.980). This means that with 95% confidence, it would not be unusual if the true geometric mean LDL levels for the group surviving less than 5 years after enrollment was between 84.8% and 98.0% as large as the true geometric mean LDL levels for the group surviving greater than 5 years after enrollment. The test for the equality of geometric means produced a statistically significant two-sided p-value of 0.0128, meaning that we can with high confidence state that this ratio indicates an association between LDL and 5-year mortality.

5. We evaluated whether or not there is an association between LDL and 5-year all-cause mortality by comparing the probability of all-cause death within 5 years after enrollment, which we will define as the risk, between groups defined by patients who have high LDL and low LDL. To compare risk of 5-year mortality between these groups, we calculated a standardized risk difference. This risk difference will be the risk of 5-year mortality for patients with high LDL subtracted from the risk of 5-year mortality for patients with low LDL.

We found that the risk of 5-year mortality for patients with high LDL was 0.137 and the risk of 5-year mortality for patients with low LDL was 0.170. We then calculated a point estimate for the risk difference of -0.033, and produced a 95% confidence interval for this difference of (-0.102, 0.036). This means that with 95% confidence, it would not be unusual if the true risk of 5-year mortality for patients with high LDL was between 0.102 lower than and 0.036 higher than the true risk of 5-year mortality for patients with low LDL. The test for equality produced a statistically insignificant p-value of 0.3753, which means that we cannot with high confidence reject the null hypothesis that there is no association between LDL levels and 5-year mortality based on this risk difference.

6. We evaluated whether or not there is an association between LDL and 5-year all-cause mortality by comparing the odds of all-cause death within 5 years after enrollment between groups defined by patients who have high LDL and low LDL. We compared these odds by calculating an odds ratio, which is equal to the odds of all-cause death within 5 years after enrollment for patients with high LDL divided by the odds of all-cause death within 5 years after enrollment for patients with low LDL.

We found that the odds of 5-year mortality for patients with high LDL was 0.158 and that the odds of 5-year mortality for patients with low LDL was 0.205. We then calculated a point estimate for the odds ratio of 0.774, with an exact 95% confidence interval for the odds ratio of (0.409, 1.387). This means that, with 95% confidence, it would not be unusual if the odds of 5-year mortality for patients with high LDL was between 40.9% and 138.7% as large as the odds of 5-year mortality for patients with low LDL. The test for equality produced a statistically insignificant p-value of 0.3753, which means that we cannot with high confidence reject the null hypothesis that there is no association between LDL levels and 5-year mortality based on this odds ratio.

7. In order to determine whether or not there is an association between LDL levels and 5-year all-cause mortality by comparing the instantaneous risk of all-cause death within 5 years after enrollment between groups defined by patients who have high LDL and low LDL, we performed a log-rank test to compare the overall survival experience of those with high and low LDL. This comparison of the overall survival experience takes into account how close the Kaplan-Meier survival curves are for each group. We also found a 95% confidence intervals for survival probability for the high LDL group and for the low LDL group at each of 1, 2, 3, 4, and 5 years after enrollment, and compared these confidence intervals to see if they overlapped.

The log-rank test gave a statistically insignificant p-value of 0.2664, so we cannot reject the null hypothesis that the overall survival experience for patients with high LDL is equivalent to that of patients with low LDL. To further support our inability to reject this null hypothesis, we can see that at each of 1, 2, 3, 4, and 5 years after enrollment, the 95% confidence interval for survival probability in the high LDL group overlap with that of the low LDL group. We note, however, that the fact that these confidence intervals is not sufficient enough to fail to reject the null hypothesis without the statistically insignificant p-value given above.



8. We would have preferred a priori the method described in question 7, comparing the instantaneous risk of death between groups defined by high LDL and low LDL, in order to answer the question about an association between mortality and serum LDL. The first reason for this choice is that this analysis seeks to describe the more clinically relevant pathway which treats 5-year mortality as the response and serum LDL as the predictor of interest, which would help address whether high LDL or low LDL is associated with higher mortality. Comparing the instantaneous risk of death allows us to consider the overall survival experience of patients in either group, which means that survival time need not be dichotomized at 5 years after enrollment, but can remain continuous, which prevents loss of potentially important time-trend information. The p-value derived from the log-rank test is useful, but has its limits since it does not accurately detect when survival curves cross, and thus is overly conservative. However, finding 95% confidence intervals for the two survival curves and comparing these confidence intervals to see whether or not they overlap at clinically relevant times (such as 1, 2, 3, 4, and 5 years after enrollment) can be useful in supplementing the conclusion drawn by the log-rank test.