1. It is valid to dichotomize the time to death according to death within 5 years of study enrollment or death after 5 years because study participants who did not die, have a minimum observation time of 5.00 years, thus there was no right censoring within 5 years among participants who did not die. We can therefore say that all of the censoring within 5 years was among participants who did die and thus it would be appropriate to dichotomize time to death. I came to this conclusion by first generating a new variable that represented observation time in years, and then looking at the summary statistics for the new binary variable by whether participants died in 5 years or survived.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Observation time in years** | **# of Participants** | **Minimum** | **Mean** | **Maximum** |
| Death=Yes | 133 | 0.19 | 3.18 | 5.54 |
| Death=No | 602 | 5.00 | 5.33 | 5.91 |

1. I first dichotomized serum LDL into low and high based on the cutoff levels used in subsequent questions in this homework. I then looked at the descriptive statistics for age, gender, weight, smoking history, stroke, CHF, and CHD by whether participants had low or high LDL levels. There are 618 participants who are considered to have low serum LDL and 117 participants with high serum LDL. Mean age, gender, smoking history and CHD were all relatively similar among the LDL groups. Thirty-two percent of participants with high LDL also had evidence of a stroke, whereas 22% of participants with low LDL had evidence of a stroke. Additionally, 6% of the low LDL participants had CHF, whereas only 3% of high LDL participants had CHF. Mean weight was slightly higher among participants with high LDL, however the medians were the same. Seventeen percent of participants with low LDL died within 5 years, whereas 14% of participants with high LDL died within 5 years.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variable | Number | Mean | Std. Dev | Minimum | 25% | Median | 75% | Maximum |
| **Low LDL (≤160 mg/dL)** | Age | 618 | 74.51 | 5.39 | 65 | 71 | 73 | 78 | 99 |
| Male | 618 | 0.51 | - | - | - | - | - | - |
| Weight (lbs) | 618 | 159.36 | 30.78 | 86 | 138 | 158 | 178 | 264 |
| Smoking History | 618 | 19.88 | 27.62 | 0 | 0 | 7 | 34 | 240 |
| CHF | 618 | 0.06 | 0.24 | - | - | - | - | - |
| Stroke | 618 | 0.22 | 0.60 | 0 | 0 | 0 | 0 | 2 |
| Death in 5 yrs | 618 | 0.17 | - | - | - | - | - | - |
| CHD | 618 | 0.33 | 0.68 | 0 | 0 | 0 | 0 | 2 |
| **High LDL (> 160 mg/dL)** | Age | 117 | 74.84 | 5.78 | 65 | 70 | 74 | 78 | 94 |
| Male | 117 | 0.44 | - | - | - | - | - | - |
| Weight (lbs) | 117 | 163.09 | 30.45 | 74 | 143 | 158 | 182 | 257 |
| Smoking History | 116 | 18.09 | 24.26 | 0 | 0 | 4 | 31 | 102 |
| CHF | 117 | 0.03 | 0.18 | - | - | - | - | - |
| Stroke | 117 | 0.32 | 0.71 | 0 | 0 | 0 | 0 | 2 |
| Death in 5 years | 117 | 0.14 | - | - | - | - | - | - |
| CHD | 117 | 0.34 | 0.70 | 0 | 0 | 0 | 0 | 2 |

1. I performed a two-sided, two-sample t test with unequal variances to determine whether there was an association between LDL levels and 5-year all-cause mortality. I did this by comparing the mean LDL levels among participants who did die within 5 years (118.70 mg/dl) compared to participants who survived after 5 years (127.20 mg/dl). The mean difference in LDL levels between the two groups is 8.5 mg/dl with a p-value of 0.0186. (Those who did die within 5 years, on average had a lower LDL levels by 8.5 mg/dl). We can therefore reject the null hypothesis that there is no difference in mean LDL levels between the two groups, at the 0.05 level. The associated 95% confidence interval of the difference in mean LDL levels between the two groups was between 1.44 and 15.56. We should proceed with caution however, in rejecting the null, because the lower bound of the 95% CI is relatively close to zero.
2. I first logged all of the LDL values and then performed a two-sided, two-sample t test with unequal variances. I then exponeniated the difference in means, as well as the corresponding 95% confidence interval. The geometric mean of LDL is 112.01 mg/dl among participants who died within 5 years. The geometric mean of LDL is 122.83 among participants who did not die within 5 years. The t-test estimated a 9.65% increase in LDL among participants who did not die within 5 years, with a p-value of 0.0128, and a 95% CI suggesting that the true geometric mean of LDL for survivors is between 1.99% and 17.87% higher than those who did not survive.
3. I first dichotomized LDL values into low and high using the above specifications and then calculated a risk difference to determine whether participants with low LDL had a lower risk of death within 5 years compared to participants with high LDL. The risk difference between the two groups is 0.033, meaning those with low LDL had a 3.3% higher risk of death within 5 years compared to those with high LDL. Based on the p-value of 0.375 and the 95% CI that includes zero (-0.102, 0.036), we cannot reject the null hypothesis that there is no association between subjects who have high serum LDL and death within 5 years.
4. I calculated an odds ratio by using logistic regression and then verified it by using the cs command in STATA. Participants with low LDL were 1.29 times more likely to die (or had 1.29 times the odds of dying) within 5 years compared to those with high LDL, however because the p-value is so large (p=0.375) and the 95% CI includes zero (0.73, 2.28), we cannot reject the null hypothesis that the odds of death within 5 years is higher among those participants with high serum LDL, compared to participants with low serum LDL.
5. I first stset my data in STATA by setting my time as “obstime” and failure as “death.” I then ran a cox regression using the binary LDL variable so that I could compare the two groups defined above. The instantaneous risk of death of participants who had low serum LDL was 25% higher for every 1 mg/dl difference in cholesterol level compared to participants with high serum LDL, (hazard ratio=0.75). We cannot reject the null hypothesis that high serum LDL participants have a higher instantaneous risk of death, however, because the p-value is 0.270 and the 95% CI includes zero (0.45, 1.25).
6. I would prefer to first perform a test where the variable LDL remains continuous, because there is a loss of information when dichotomizing it. This leaves either of the two-sample two-sided t-tests with unequal variances. I ultimately would have chosen the two-sided, two-sample t-test using geometric means because it has the largest Z value (as shown below in the table below). The Z statistics tells us how many standard errors we are away from the null hypothesis of no association, so the larger the value of the Z statistic; the more likely it is we can reject the null.

|  |  |
| --- | --- |
| **Test** | **Z or T Statistics** |
| T-test (death as binary) | 2.39 |
| T-test using geometric mean (death as binary) | 2.52 |
| Risk Difference (both binary) | 2.50 |
| Logistic (both binary) | 0.88 |
| Cox Regression (LDL as binary) | -1.10 |