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 5 years of study enrollment or death after 5 years. Why is this valid? Provide descriptive statistics that support your answer.

Our data come from a cross-sectional study, and Table 1 shows that no censoring occurred before 5 years. Specifically, the minimum follow-up time for patients subject to administrative censoring was 1827 days (5.001 years). Since all of the patients who were censored after 5 years also must have died after 5 years, there are no patients for whom we cannot determine whether time to death was less than or greater than five years. Thus, it is valid to use 5 years as the cut-off for dichotomizing time to death and thereafter analyze the data using methods that disregard censoring.

Table 1. Descriptive statistics for the distribution of observation time (in years) to death or censoring by patient outcome. Statistics provided include the number of observations (N), the mean, the median (Med), the 25th and 75th percentiles (interquartile range, or IQR), the minimum (Min), and the maximum (Max).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Outcome | N | Mean (SD) | Med (IQR) | (Min, Max)  |
| Death | 133 | 3.2 (1.5) | 3.5 (2.0, 4.5) | (0.2, 5.5)  |
| Censoring | 602 | 5.3 (0.3) | 5.2 (4.1, 5.7) | (5.0, 5.9) |

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 LDL and 5 year all-cause mortality in the medical literature. In attention to the two variables of primary interest, you may restrict attention to age, sex, weight, smoking history, and prior history of cardiovascular disease (coronary heart disease (CHD), congestive heart failure (CHF), and stroke.

Of the 735 participants in this study, 366 were male and 369 were female. The mean ages were very similar for persons with low serum LDL (low density lipoprotein), moderate LDL, and high LDL; these ages were 74.8, 74.4, and 74.9 years, respectively. Mean weights were also very similar, at 159.9, 159.2, and 162.7 pounds for the same groups as above. There was a slightly higher proportion of smokers among those with low LDL (17.0%) compared to those with moderate (12.6%) and high (12.1%) LDL.

The proportion of people suffering from myocardial infarctions was slightly lower in the group with low LDL (9.7%) compared to those with moderate and high LDL (13.2% and 12.1%, respectively). Similarly, the proportion of people suffering from strokes was higher in the group with high LDL levels (13.1%) compared to the groups with low and moderate LDL levels (9.1% and 9.7%, respectively). However, other adverse events were rarer in the groups with higher LDL levels in our sample. Incidence of congestive heart failure decreased with increasing LDL levels, from 9.1% (low LDL) to 4.9% (moderate) and then 2.8% (high LDL). Similarly, the proportion of people who died within five years decreased as LDL increased, with a proportion of 20.0% among people with low LDL, 15.9% among those with moderate LDL, and 13.1% among people with high LDL.

|  |
| --- |
| Table 2. Descriptive statistics for characteristics of study subjects, stratified by serum LDL (low density lipoprotein) levels at the time of the subject’s MRI. According to the Mayo Clinic’s guidelines, serum LDL levels below 100 mg/dL are ideal for people at risk of heart disease, levels between 100 and 160 mg/dL are ideal or slightly high, and levels above 160 mg/dL are high.  |
| **Serum LDL (mg/dL)** | **<100** | **100-160** | **≥160** | **Total** |
| **N** | 165 | 453 | 107 | 735 |
| **Male, n (%)**  | 95 (57.6) | 220 (48.6) | 45 (42.1) | 366 (49.8) |
| **Age (yrs), mean [SD]** | 74.8 [5.5] | 74.4 [5.4] | 74.9 [5.8] | 74.6 [5.5] |
| **Weight (lbs), mean [SD]** | 159.9 [31.5] | 159.2 [30.6] | 162.7 [30.7] | 159.9 [30.7] |
| **Smoking history** |  |  |  |  |
| **Never smokers, n (%)**  | 79 (47.9) | 188 (41.5) | 49 (45.8) | 319 (43.4) |
| **Former smokers, n (%)**  | 58 (35.1) | 206 (45.5) | 45 (42.1) | 314 (42.7) |
| **Current smokers, n (%)**  | 28 (17.0) | 57 (12.6) | 13 (12.1) | 99 (13.5) |
| **Coronary heart disease** |  |  |  |  |
| **None, n (%)** | 135 (81.8) | 353 (77.9) | 86 (80.4) | 580 (78.9) |
| **Angina, n (%)** | 14 (8.5) | 40 (8.8) | 8 (7.5) | 64 (8.7) |
| **Mycoardial infarction, n (%)** | 16 (9.7) | 60 (13.2) | 13 (12.1) | 91 (12.4) |
| **Congestive heart failure, n (%)**  | 15 (9.1) | 22 (4.9) | 3 (2.8) | 41 (5.6) |
| **Stroke**  |  |  |  |  |
| **None, n (%)** | 142 (86.1) | 399 (88.1) | 87 (81.3) | 636 (86.5) |
| **Transient ischemic attack, n (%)**  | 8 (4.8) | 10 (2.2) | 6 (5.6) | 24 (3.3) |
| **Stroke, n (%)**  | 15 (9.1) | 44 (9.7) | 14 (13.1) | 75 (10.2) |
| **Death within 5 years, n (%)**  | 33 (20.0) | 72 (15.9) | 14 (13.1) | 121 (16.5) |

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing mean LDL values across groups defined by vital status at 5 years.

Methods: I performed a two-sample t-test allowing for unequal variances to compare the arithmetic mean serum LDL values between patients who died within five years of their MRI test and patients who lived at least five years after their MRI test.

Inference: Our null hypothesis is H0: μ1 - μ2 =0, where μ1 refers to the mean LDL among patients who lived at least five years and μ2 refers to the mean LDL among patients who died within five years of their MRI test. The alternative hypothesis is μ1 - μ2≠0. In our data, μ1 = 127.2 and μ2 = 118.7, so our estimated difference is 8.5 mg/dL. A 95% confidence interval for the difference is (1.4, 15.6), with an associated p-value of 0.0186. Thus, we can reject our null hypothesis with 95% confidence and conclude that mean serum LDL differs between those who die within 5 years of their MRI and those who live at least 5 years.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing geometric mean LDL values across groups defined by vital status at 5 years.

Methods: To compare the geometric mean serum LDL values between patients who died within 5 years of their MRI test and patients who lived at least 5 years after their MRI, I performed a two-sample t-test allowing for unequal variance on the log-transformed data, then exponentiated the estimates and confidence interval.

Inference: Our null hypothesis is H0: μ1 / μ2 =1, where μ1 refers to the geometric mean LDL value among patients who lived at least five years and μ2 refers to the geometric mean LDL value among patients who died within five years of their MRI test. The alternative hypothesis is μ1 / μ2 ≠ 1. The estimated difference between geometric means is 0.092 mg/dL. A 95% confidence interval for the difference is (0.85, 0.98), with an associated p-value of 0.013. Thus, we can reject our null hypothesis with 95% confidence and again conclude that mean serum LDL differs between those who die within 5 years of their MRI and those who live at least 5 years.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the probability of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL ≥ 160 mg/dL).

Methods: I performed a chi-square test with one degree of freedom to test to test the null hypothesis that high serum LDL levels and 5 year all-cause mortality are independent, with the alternative hypothesis that LDL levels and mortality are not independent. A binomial Wald confidence interval was created for the risk difference comparing the probability of death within five years between those with high serum LDL and those with low to moderate serum LDL.

Inference: Based on the chi-square test, we obtain a p-value of 0.375, which is insufficient evidence to conclude that serum LDL levels are not independent of 5 year all-cause mortality. The difference in probability of death within 5 years was 3.3%, with patients with low to moderate LDL having a higher probability of death. This estimate is consistent with a true difference in proportions between 3.6% lower and 10.2% higher among those with low to moderate serum LDL values.

1. Perform a statistical analysis evaluating an association between serum LDL and 5 year all-cause mortality by comparing the odds of death within 5 years across groups defined by whether the subjects have high serum LDL (“high” = LDL ≥ 160 mg/dL).

Methods: I created a Woolf confidence interval for the odds ratio comparing the odds of death within 5 years for those who have high versus low to moderate serum LDL values. The null hypothesis is that the odds of death are equal between the group with high LDL and the group with low or moderate LDL, that is, that the odds ratio is 1. The alternative hypothesis is that the odds ratio is not equal to 1, and the odds of death differ between the group with high LDL levels and the group without high LDL levels. The p-value was calculated using a chi-square test, as in Question 5.

Inference: The estimate of the odds ratio is 1.29. This estimate is consistent with a true value between 0.732 and 2.279. Since the confidence interval includes one, the data are consistent with a true ratio of 1, that is, with equal odds of death regardless of serum LDL levels. Thus, based on this evidence, we cannot reject the null hypothesis and conclude that the odds of death within five years differ between groups defined by serum LDL levels. This is also consistent with the chi-square p-value, which is 0.037, as mentioned above.

1. Perform a statistical analysis evaluating an association between serum LDL 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 LDL (“high” = LDL ≥ 160 mg/dL).

Methods: A Kaplan-Meier failure curve with 95% confidence intervals was created showing time until death stratified into groups defined by whether subjects have high (≥160 mg/dL) serum LDL levels.

Inference: Figure 1 shows Kaplan-Meier estimates of the probability of death by a given time point in the sample. In general, the group with LDL <160 mg/dL had a higher probability of dying over a time period. However, the cumulative hazard functions for the two groups do not differ by much, as both 95% confidence intervals encompass both hazard curves. After three years, 92% of the group with low LDL at the time of their MRI exam remained alive, compared to 94% in the group with high LDL, as noted in Table 3. After five years, 83% of the group with low LDL remained alive, compared to 86% of the group with high LDL.



Figure 1. Kaplan-Meier failure functions for time to death stratified into groups based on serum LDL levels, where high serum LDL is defined as ≥160 mg/dL.

|  |
| --- |
| Table 3. Descriptive statistics for the distribution of time (in days) to death by LDL category. Statistics provided include the number of observations (N) and the number of observed events (Ev), the minimum (Min) and censored maximum (Max) survival times, and the probability of surviving until 3 or 5 years after the MRI exam.  |
|  | **N (Ev)**  | **(Min, Max)** | **3 yr Surv Prob (CI)** | **5 yr Surv Prob (CI)** |
| **LDL <160 mg/dL** | 618 (116) | (68, >2158) | 0.92 (0.89, 0.94) | 0.83 (0.80, 0.86) |
| **LDL >160 mg/dL**  | 107 (15) | (415, >2158) | 0.94 (0.88, 0.97) | 0.86 (0.79, 0.91) |
| **Total** | 725 (131) | (68, >2158) | 0.92 (0.90, 0.94) | 0.84 (0.81, 0.86) |

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 LDL? Why?

*A priori*, I would have preferred to perform a Kaplan-Meier survival analysis. In all of the analyses we have learned thus far, we must choose to dichotomize the time to death (e.g., questions 3 and 4 above), dichotomize LDL values (e.g., question 7), or both (e.g., questions 5 and 6). Dichotomization necessarily discards data, so the question at hand is which data we prefer to discard. The question we are most interested in is whether lower (or higher) LDL values are associated with longer survival, so the more natural analyses will use LDL values as the grouping variable or predictor of interest. This requires that LDL values be dichotomized. Also, considering low (<160 mg/dL) versus high (≥160 mg/dL) LDL is a dichotomization of clinical relevance, since these are the cutoff values used in clinical practice. In contrast, five-year survival is not chosen because of any particular clinical relevance; we are interested in overall long-term survival, and five-year survival would simply be a measurable stand-in for our true variable of interest.

Knowing that we will dichotomize LDL, we must still choose whether to also dichotomize time to death and analyze proportions or whether to perform the Kaplan-Meier survival analysis. Given this choice, I would prefer the lesser reduction of the data, so I would have chosen to perform a Kaplan-Meier survival analysis, which does not require us to dichotomize the time to death. This also allows us to see the consistency or pattern of association over time.

I do note that regression analysis would not require dichotomization of either variable, but as we have not fully covered that yet, I would not have been likely to choose that analysis method.