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

The following descriptive statistics tables provide the reason why we are allowed to dichotomize the time to death as described. Within 5 years of enrolment, death indicator is all 1 (mean, min and max are 1). This means there were 121 patients died within 5 years and none of these data is censored. In fact, the first censored data is recorded just a bit after 5-year period according to the second descriptive statistics table. Since we wouldn’t be able to tell the exact survival time of these patients who have censored data, it’s appropriate to put them in the group of “survive longer than 5 years” as opposed to the group “survive less than 5 years”.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Death Indicator |  |  |  |  |
| Time | N | mean | sd | min | max |
| Within 5 years | 121 | 1 | 0 | 1 | 1 |
| After 5 years | 614 | 0.0195 | 0.139 | 0 | 1 |
| Total | 735 | 0.1810 | 0.385 | 0 | 1 |

|  |  |
| --- | --- |
|  | Time |
| Death | min |
| 0 | 5.005 |
| 1 | 0.186 |

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.

The following table provides descriptive statistics categorized by time to death recorded (less than 5 years vs. longer than 5 years). There were 121 patients die within 5 years of measurement and 614 patients survived at least 5 years. Measurement of how many pack of cigarette patient smokes per day for 1 year is missing for 1 person in the < 5 years group. It seems like the proportion of male is greater in the < 5 years group. Also, more patients in this group tend to smoke, have prior history of CHD, CHF and have either a transient ischemic attack or a stroke prior to MRI.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Time to death < 5 yrs. N; Mean; SD; Min; Mdn; Maxn (%) | Time to death > 5 yrs. N; Mean; SD; Min; Mdn; Maxn (%) | All patients. N; Mean; SD; Min; Mdn; Maxn (%) |
| Age (years) | 121; 76.48; 6.170; 67; 75; 91 | 614; 74.19; 5.222; 65; 73; 99 | 735; 74.57; 5.451; 65; 74; 99 |
| # of male patients  | 78 (64.5%) | 288 (46.9%) | 366 (49.8%) |
| Weight (lbs) | 121; 159.12; 32.789; 96; 154; 264 | 614; 160.11; 30.347; 74; 158.75; 258 | 735; 159.95; 30.741; 74; 158; 264 |
| Smoke (pack/day for 1 year) | 120; 28.05; 36.042; 0; 18.375; 240 | 614; 17.950; 24.695; 0; 4.35; 180 | 734; 19.60; 27.108; 0; 6.5; 240 |
| # of years since quitting smoking | 121; 10.71; 14.445; 0; 0; 56 | 614; 9.45; 14.034; 0; 0; 56 | 735; 9.66; 14.100; 0; 0; 56 |
| CHF | 17 (14.0%) | 24 (3.9%) | 41 (5.6%) |
| CHD | .6198347; NA; 0; 0; 2 | .2785016; NA; 0; 0; 2 | .3346939;NA; 0;0; 2 |
| Stroke | 121; 0.52; NA; 0; 0; 2 | 614; 0.181; NA; 0; 0; 2 | 735; 0.24; NA; 0; 0; 2 |

\*NA means that statistics is meaningless for that variable.

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
| 0 | 606 | 127.198 | 1.338 | 32.929 | 124.571 | 129.825 |
| 1 | 119 | 118.698 | 3.315 | 36.157 | 112.134 | 125.261 |
| Combined | 725 | 125.803 | 1.248 | 33.602 | 123.353 | 128.253 |
| Difference |  | 8.501 | 3.574 |  | 1.441 | 15.560 |
| Ho: difference = 0  | Ha: difference ≠ 0 |
| t-statistic = 2.3783 |
| p-value (2-sided t-test) = 0.0186 | p-value (1-side) = 0.0093 |

The above table provides the result of a 2-sample t-test testing the if there is a significant difference in mean ldl between patients who died within 5 years (group 1) and who survived (group 0). The estimate of mean ldl is 127.2 mg/dL for the group without death observed and 118.7 mg/dL for the group with death observed. Thus, survived group has a mean ldl level that is 8.5 mg/dL higher than the ldl level in death group. The difference is significantly different from 0 (P-value < 0.05) at 95% significance level. The 95% confidence interval shows that our estimate of the difference would be reasonable if the true difference is between 1.441 mg/dL and 15.56 mg/dL. Therefore, we can reject the null hypothesis of no difference of mean ldl between patients who died within 5 years and those who survived after 5 years of study in favor of alternative hypothesis that is patients who survived after 5 years tend to have a higher ldl level.

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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Group | Obs | Mean | Std. Err. | Std. Dev. | [95% Conf. | Interval] |
| 0 | 606 | 4.811 | 0.011 | 0.270 | 4.789 | 4.832 |
| 1 | 119 | 4.719 | 0.035 | 0.380 | 4.650 | 4.788 |
| Combined | 725 | 4.796 | 0.011 | 0.293 | 4.774 | 4.817 |
| Difference |  | 0.092 | 0.037 |  | 0.020 | 0.164 |
| Geometric mean for group 0: 122.825 |
| Geometric mean for group 1: 112.011 |
| Ratio of geometric mean of group 0 and 1: 9.65% |
| 95% C.I. of ratio of geometric mean of group 0 and 1: [2.01%, 17.87%] |
| p-value (2-side) : 0.0128 | p-value (1-side) : 0.0064 |

The estimate of geometric mean of ldl is 112.01 mg/dL for those patients died within 5 years of study (Group 1) and 122.83 for those who survived at least 5 years (Group 0). Compare the estimate of ratio of geometric means of two groups, we get geometric mean of ldl for group 0 is 9.65% higher than the geometric mean for group 1. The difference is significantly different from 0 at 95% significant level (p-value < 0.05). 95% confidence interval shows that the estimate of the ratio of group 0 and group 1 is reasonable if group 0 patients have a geometric mean that is truly higher than group 1 patients from 2.01% to 17.87%. Hence, we can reject null hypothesis that assumes no association of mean geometric ldl between survivor and nonsurvior within 5 years in favor of alternative hypothesis that assumes higher geometric mean of ldl among the patients who survived longer than 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).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Mean | Std. Err. | [95% Conf. | Interval] |
| 0 | 0.170 | 0.015 | 0.140 | 0.200 |
| 1 | 0.131 | 0.033 | 0.067 | 0.195 |
| Difference | 0.039 | 0.036 | -0.031 | 0.109 |
| Z-statistics = 1.0072 |
| p-value (2-sided) : 0.3139 |
|  | High LDL |  |  |
|  | Exposed | Unexposed | Total |
| Death |  |  |  |
| Cases | 14 | 105 | 119 |
| Control | 93 | 513 | 606 |
| Total | 107 | 618 | 725 |
| Risk | 0.131 | 0.170 | 0.164 |
| Risk difference: -.0390 |
| 95% C.I: [-0.1095 0.031] |
| Chi-square statistic: 1.01 | P-value: 0.3139 |

The tables above provide the result from a chi-square test. Null hypothesis is that the probability of death within 5 years has no difference across the groups defined by high and low LDL. Alternative hypothesis is that there’s a significant difference of probability of death with 5 years across the groups. Our estimate of probability of death with low ldl is 17% based on the fact that 105 out of 618 people in the low ldl group died within 5 years. Similarly, the estimate of probability of death with high ldl is 13% within 5 years of study. Thus, probability of death of **low** ldl group is 3.9% **higher** than probability of death of **high** ldl group. According to the 95% confidence interval, this observed difference is reasonable if the true probability difference (probability of death in high ldl group – probability in low ldl group) is from -10.95% to 3.1%. However, this difference is not significant different from 0 (p-value > 0.05) at 95% significance level. Therefore, we cannot reject the null hypothesis that assumes the probability of death of low ldl groups is equal to the probability of death of high ldl at 95% significance level.

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).

|  |  |
| --- | --- |
| Odds ratio | 0.735 |
| [95% Conf. Interval] | [.373, 1.36] |
| Chi-square statistic: 1.01 | P-value: 0.3139 |

 The table above provides the result of a case-control odds ratio. We expect the odds of death happened within 5 years for patients with high ldl to be 26.5% lower (odd ratio is 73.5%) than the odds of death for patients in low ldl group. Also, the 95% confidence interval suggests that the observed odds ratio would be reasonable if the true odd of death of high ldl group is within the range from 62.7% lower to 36.0% higher in terms of the odds of death of low ldl group. However, the odds ratio is not significantly different from 1 (p-value > 0.05) at 95% significance level. That is saying we cannot reject null hypothesis which assumes there is no difference of odds between two groups (no association between ldl and mortality).

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).

|  |  |  |
| --- | --- | --- |
|  | Events |  |
| High LDL | observed | expected |
| 0 | 116 | 111.01 |
| 1 | 15 | 19.99 |
| Total | 131 | 131 |
| Chi-square statistics: 1.47 |
| P-value: 0.2249 |

The table above provides the result of a logrank test. Null hypothesis is that there’s no difference of hazard (instantaneous death risk) between 2 groups (Hazard ratio = 1). According to the test result, the hazard ratio is not significantly different from 1 (p-value > 0.05) at 95% significance level. That is saying we cannot reject null hypothesis which assumes there is no difference of instantaneous risk between two groups (no association between ldl and mortality).

**\*We can also use cox regression for this question and it will give us more information (point estimator and confidence interval).**

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?

The analysis in question 3 and 4 are to study the ldl level conditioned on if death happened within 5 years or not whereas the rest of the analysis (5-7) are based on if patients are in the low ldl group or high ldl group. Since we were to test if there’s an effect of ldl on mortality, I think the latter study design is better because people are often more interested in the probability of case (death) if exposed/unexposed (in our case, it’s high/low ldl) rather than knowing if exposed given case. Predicting future seems to be more important than studying the past in this situation. According to the scientific question we are interested in this study, analysis in question 5, that is to compare the probability of death within 5 years across groups defined by whether the subjects have high serum LDL is reasonable. The insignificant result in this study showed that high ldl might not be an effect on mortality. It can be a useful evidence to be provided to the people who are concerning the chance of death after being diagnosed to have high ldl. However, we lost information of time to death by splitting the time to death into 2 groups. Therefore, study design in question 7 might be the most appropriate one to use for this dataset since it counts the censored data. The drawback of that is the logrank test provided us nothing but the p-value. Cox regression model would be more informative.

Method of design in question 6 is not preferred since odds ratio is not the interest of study.