Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2014

Homework #1 Key
January 13, 2014

***Instructions for grading*: Prior to the answer for each problem, I provide the maximum points to be given for each problem, and the way that points should be distributed. Please insert comments on to the document indicating the points you have awarded for the problem, commenting on any reasons points were deducted. My answer to each question is provided in boldface type. In giving the answers, I sometimes provide alternative approaches in order that you can assess whether the numbers match up. I also provide some discussion of the choices or some additional material that I did not really expect to be provided in the answer. This additional information is provided in** *normal type.*

Written problems: To be submitted as a MS-Word compatible file to the class Catalyst drop box by 9:30 am on Monday, January 13, 2014. See the instructions for peer grading of the homework that are posted on the web pages.

> *On this (as all homeworks) Stata / R code and unedited Stata / R output is **TOTALLY** unacceptable. Instead, prepare a table of statistics gleaned from the Stata output. The table should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits. (I am interested in how statistics are used to answer the scientific question.)*
>
> ***In all problems requesting "statistical analyses" (either descriptive or inferential), you should present both***
>
> - ***Methods:* A brief sentence or paragraph describing the statistical methods you used. This should be using wording suitable for a scientific journal, though it might be a little more detailed. A reader should be able to reproduce your analysis. DO NOT PROVIDE Stata OR R CODE.**
> - ***Inference*: A paragraph providing full statistical inference in answer to the question. Please see the supplementary document relating to "Reporting Associations" for details.**
>
> *Keys to past homeworks from quarters that I taught Biost 517 (e.g. HW #8 from 2012) or Biost 518 (e.g., HW #3 from 2008) or Biost 536 (e.g. HW #3 from 2013) might be consulted for the presentation of inferential results. Note that the requirement to provide a paragraph describing your statistical methods is new this year, and thus past keys do not give explicit examples of a separate paragraph. However, many past keys provide this information as an introductory sentence.*

All questions relate to associations between death from any cause and serum low density lipoprotein (LDL) levels in a population of generally healthy elderly subjects in four U.S. communities. This homework uses the subset of information that was collected to examine MRI changes in the brain. The data can be found on the class web page (follow the link to Datasets) in the file labeled mri.txt. Documentation is in the file mri.pdf. The data is in free-field format, and can be read into Stata using the following code in a .do file.

```
infile ptid mridate age male race weight height packyrs yrsquit alcoh ///
    physact chf chd stroke diabetes genhlth ldl alb crt plt sbp aai ///
    fev dsst atrophy whgrd numinf volinf obstime death ///
    using http://www.emersonstatistics.com/datasets/mri.txt
```

Note that the first line of the text file contains the variable names, and will thus be converted to missing values. Similarly, there is some missing data recorded as 'NA', and those, too, will be converted to missing values. If you do not want to see all the warning messages, you can use the "`quietly`" prefix. You may want to go ahead and drop the first case using "`drop in 1`", because it is just missing values.

Recommendations for risk of cardiovascular disease according to serum LDL (low density lipoprotein) levels are as follows (taken from the Mayo Clinic website):

Below 70 mg/dL	Ideal for people at very high risk of heart disease
Below 100 mg/dL	Ideal for people at risk of heart disease
100-129 mg/dL	Near ideal
130-159 mg/dL	Borderline high
160-189 mg/dL	High
190 mg/dL and above	Very high

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.

***Instructions for grading*: This problem is worth 5 points. To get any credit, the answer must note the minimum time of follow-up for a censored observation.**

Ans: The minimum time of follow-up among censored observations is 1,827 days, or just over 5 years. Hence the vital status of every individual is known at 5 years. *(This is about the only reason that it is useful to look at sample descriptive statistics on a censored variable. All other uses of descriptive statistics should use Kaplan-Meier estimates.)*

2. 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 addition 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.

***Instructions for grading*: This problem is worth 10 points. Assign 4 points to general table layout and labeling of columns, rows, and descriptive statistics, assign 3 points to choice of descriptive statistic, and assign 3 points to the discussion of the finding. The ultimate score should be based on your ability to understand what is presented. Columns should typically correspond to the groups being compared, rows should correspond to the individual variables. Columns should be clearly labeled in scientific terms. Similarly, rows should be labeled with the corresponding variable to which the descriptive statistics apply. The names of the variables should be in English, not any nonstandard abbreviations used in computer coding. (LDL is a standard abbreviation, though it does not hurt to have it defined somewhere.) Units for the variables should be made clear. It should be clear which descriptive statistics are presented in the table. The descriptive statistics presented for continuous random variables should include, at a minimum, the sample size in each group, the number of cases with missing data (this could be a column, or just a footnote), the mean, and the standard deviation. The minimum and maximum might also be included. While the median and/or interquartile range might also be included, I note that they do not help us judge confounding as well. Descriptive statistics should generally include three significant digits, though some variation is possible due to the values in the different columns. The student should provide some general comments on what the descriptive statistics tell us relative to the types of patients and the possibility of confounding.**

Ans: *In choosing how to answer this question, we should consider the goals of descriptive statistics. In lecture, I present 5 reasons: 1) detecting errors, 2) describing materials and methods, 3) assessing validity of assumptions, 4) straightforward estimates addressing the primary question of interest, and 5) exploratory analyses. For this problem, the major role of the descriptive statistics should be the second (we want to know the types of patients used in our analysis) and the third (we would like to assess any potential confounding). In the same table we can also address the fourth (descriptive statistics that relate to any association between mortality and serum LDL). Journal editors do not like extensive tables, so we must try to economize somewhat as we try to address the three different goals of the descriptive statistics.*

In terms of describing the types of patients used in the study, we will want to know the number of subjects (an important clue to statistical precision and generalizability), any patterns of missing data (an important clue to credibility of analyses), some measure(s) of central tendency (mean, median, geometric mean), and some measure(s) of spread (standard deviation, range, interquartile range).For the "materials and methods", many choices make sense. Means and SD tend to be standard, unless the data is extremely skewed, in which case the median might be more indicative of "central tendency". For certain variables, reporting the geometric mean might be standard, but it is rare to report it in a table of descriptive statistics without really good reason. Knowing the minimum and maximum is nice, when it has not been determined from the study design. For categorical data, we report frequencies. And we sometimes would divide a continuous variable into scientifically important categories and report frequencies within each category.

For the purposes of assessing the possibility of confounding, we should consider the properties of a confounder: 1) a confounder must be causally associated with the outcome variable, independently of the predictor of interest (i.e., not in the causal pathway of interest), <u>and</u> 2) the confounder must be associated with the predictor of interest (POI) in the sample. We can fairly easily use our presentation of descriptive statistics to help address confounding by providing those descriptive statistics in <u>columns</u> (it is easiest for us to make comparisons across columns, rather than rows) defined by either the outcome variable or defined by the predictor of interest. Factors that should be considered when choosing between these two options include:

- *The sampling scheme. If we have constrained the sample size within any groups, the columns should be based on that sampling. Hence, in a case-control study, the columns have to be defined by disease status. In a cohort study in which the sample sizes for the exposure group were set by design, the columns have to be defined by exposure. In cross-sectional sampling or in a cohort study in which only the total sample size was constrained, we can choose either approach.*
- *The greatest value when trying to assess confounding. A confounder has to be causally associated with the outcome variable (at least to the best of our knowledge). Hence, we probably already have a pretty good idea about the associations that we will see between the outcome and the other variables (besides the POI). So the value added by making columns defined by the outcome variable is perhaps less than that added when making columns defined by the POI: Confounders have to be associated with the POI in the sample, and it is possible that associations that exist in the population do not exist in the sample (perhaps by study design) and vice versa (no association exists in the population, but we got unlucky in our sampling). Hence, I have a definite preference for displaying descriptive statistics within columns defined by the POI, when all other things are equal.*
- *The need to dichotomize the variables. In order to display descriptive statistics within groups, we may have dichotomize or trichotomize a continuous variables. When scientifically relevant thresholds are known (e.g., I provided some information from the Mayo Clinic regarding thresholds used for LDL), this is not such an issue. But if one variable is already dichotomized and the other one not, that might push me in that direction.*

All things considered, in this case I prefer to divide the sample into groups based on LDL. I decided to provide three categories just to be able to get some idea of consistency of trends. I note that I choose the intervals based on prior scientific knowledge. It would be misleading to start exploring data and decide on intervals that accentuate differences: such a process is quite likely to introduce bias, and thus not be descriptive.

Methods: Indicator variables were created for prior history of myocardial infarction (MI), prior history of angina with no history of MI, prior history of stroke (cerebrovascular accident or CVA), prior history of transient ischemic attacks (TIA) without stroke, prior history of any cardiovascular disease (angina, MI, CVA, TIA, or congestive heart failure (CHF), and death within 5 years of study enrollment (no subjects were censored during that period of observation). Descriptive statistics are presented within groups defined by serum LDL measurements (less than or equal to 129 mg/dL, between 130 and 159 mg/dL inclusive, and greater than or equal to 160 mg/dL), as well as in the entire sample. Within each group defined by serum LDL level, for continuous variables (age, weight, pack years of smoking) we include the mean, standard deviation, minimum and maximum. For binary variables (sex and indicators of prior history of angina w/o MI, MI, TIA w/o stroke, stroke, CHF, CVD, or death) we present percentages.

Results: Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are missing data on serum low density lipoprotein (LDL). Those subjects are omitted from all analyses, but it should be remembered that we can not assess the impact that such omissions might have on the generalizability of our results. None of the 725 subjects were missing data on any other variables of interest for this analysis.

Of the 725 subjects with available measurements, 393 had serum LDL measurements less than or equal to 129 mg/dL, 225 had measurements between 130 mg/dL and 159 mg/dL inclusive, and 107 had measurements greater than or equal to 160 mg/dL. The following table presents descriptive statistics within these groups. Subjects having serum LDL in the lowest interval were more likely to be male than in other intervals. No consistent trend was seen across groups in age, weight, smoking history, or prior cardiovascular disease. In particular, there was marginally higher prevalence of prior cardiovascular disease in both the lowest and highest LDL groups (32.6% and 31.8%, respectively) compared to those with intermediate values (25.8%). Subjects with the lowest levels of serum LDL appeared to have a higher mortality rate: 19.3% of subjects with LDL less than or equal to 129 mg/dL died within 5 years compared to about 13% in subjects with higher serum LDL at study entry.

	Serum Low Density Lipoprotein (LDL)			
	≤ 129 mg/dL (n=393)	**130 - 159 mg/dL (n=225)**	**≥ 160 mg/dL (n=107)**	**Any Level (n=725)**
Male (%)	55.5%	43.1%	42.1%	49.7%
Age (yrs)[1]	74.7 (5.25; 65 - 92)	74.2 (5.62; 67 - 99)	74.9 (5.77; 65 - 94)	74.6 (5.45; 65 - 99)
Weight (lbs)[1]	160 (29.9; 86 - 264)	158 (32.3; 96 - 245)	163 (30.7; 74 - 257)	160 (30.8; 74 - 264)
Smoking (pack-years)[1]	19.8 (26.94; 0 - 180)	20.0 (28.83; 0 - 240)	18.1 (24.41; 0 - 102)	19.6 (27.16; 0 - 240)
Prior angina w/o MI (%)	10.2%	6.2%	7.5%	8.6%
Prior MI (%)	12.2%	12.4%	12.1%	12.3%
Prior TIA w/o stroke (%)	3.8%	1.3%	5.6%	3.3%
Prior Stroke (%)	9.2%	10.2%	13.1%	10.1%
Prior CHF (%)	6.6%	4.9%	2.8%	5.5%
Prior CVD (%)	32.6%	25.8%	31.8%	30.3%
Death w/in 5 years	19.3%	12.9%	13.1%	16.4%

[1] Descriptive statistics presented are the mean (standard deviation; minimum – maximum)

Below I also provide an answer based on dividing the sample according to survival for at least 5 years after study entry. I suppose I could have also divided this table into more categories, but the only reason I can really think of for preferring this table to the former table based on LDL is that the mortality data is already dichotomized for most of the analyses you are asked to do. I believe you will see that I mostly used cut and paste from the previous answer to provide this answer.

Methods: **Indicator variables were created for prior history of myocardial infarction (MI), prior history of angina with no history of MI, prior history of stroke (cerebrovascular accident or CVA), prior history of transient ischemic attacks (TIA) without stroke, prior history of any cardiovascular disease (angina, MI, CVA, TIA, or congestive heart failure (CHF), and death within 5 years of study enrollment (no subjects were censored during that period of observation). Descriptive statistics are presented within groups defined by death within 5 years, survival for 5 years post study entry, and for the entire sample. Within each group defined by vital status at 5 years, for continuous variables (age, weight, pack years of smoking, low density lipoprotein (LDL)) we include the mean, standard deviation, minimum and maximum. For binary variables (sex and indicators of prior history of angina w/o MI, MI, TIA w/o stroke, stroke, CHF, CVD, or death) we present percentages.**

Results: **Data is available on 735 subjects, however 10 of those subjects (including 2 who died within 5 years) are missing data on serum low density lipoprotein (LDL). Those subjects are omitted from all analyses, but it should be remembered that we can not assess the impact that such omissions might have on the generalizability of our results. None of the 725 subjects were missing data on any other variables of interest for this analysis.**

Of the 725 subjects with available measurements, 119 died within 5 years of study enrollment and 606 were still alive 5 years after study enrollment. The following table presents descriptive statistics within these groups. Subjects dying within 5 years were more likely to be male, tended to be older, tended toward a more extensive history of smoking, and tended to have higher prevalence of all categories of cardiovascular disease than subjects surviving for at least 5 years after study enrollment. Subjects with dying within 5 years also tended toward lower serum LDL at study enrollment: mean serum LDL was 118.7 mg/dL in those observed to die within 5 years compared to a mean serum LDL of 127.2 mg/dL in those surviving at least 5 years.

	Vital Status at 5 Years Post Study Enrollment		
	Alive at 5 Years (n=606)	**Death w/in 5 Years (n=119)**	**All Subjects (n=725)**
Male (%)	46.7%	64.7%	49.7%
Age (yrs)[1]	74.2 (5.21; 65 - 99)	76.6 (6.16; 67 - 91)	74.6 (5.45; 65 - 99)
Weight (lbs)[1]	160 (30.4; 74 - 258)	159 (33.0; 96 - 264)	160 (30.8; 74 - 264)
Serum LDL (mg/dL)[1]	127.2 (32.93; 39 - 247)	118.7 (36.16; 11 - 227)	125.8 (33.60; 11 - 247)
Smoking history (pack-years)[1]	18.0 (24.72; 0 - 180)	28.0 (36.19; 0 - 240)	19.6 (27.16; 0 - 240)
Prior angina w/o MI (%)	7.4%	14.3%	8.6%
Prior MI (%)	10.1%	23.5%	12.3%
Prior TIA w/o stroke (%)	2.8%	5.9%	3.3%
Prior Stroke (%)	7.6%	22.7%	10.1%
Prior CHF (%)	3.8%	14.3%	5.5%
Prior CVD (%)	25.1%	57.1%	30.3%

[1] Descriptive statistics presented are the mean (standard deviation; minimum – maximum)

3. 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.

***Instructions for grading*: This problem is worth 10 points. Assign 5 points to performing an appropriate analysis and describing the methods appropriately, and 5 points to reporting the association appropriately. To receive full credit for reporting the association, the answer must make clear:**

- ***the variable whose distribution is being summarized (the response variable,)***
- ***the summary measure of that distribution that is being compared across groups,***
- ***the groups that are being compared (in scientific wording),***
- ***how those groups are being compared (difference or ratio),***
- ***an estimate (and units) of the association (and for two sample problems, it is nice if point estimates of the individual groups are given when possible),***
- ***a confidence interval for the estimate of the association, and***
- ***a p value and conclusion about the association (including whether one-sided or two-sided).***

Ans: *In choosing how to answer this question, there are basically two options that would typically be considered: the t test that presumes equal variances or the t test that allows for the possibility of unequal variances. (We do actually have other methods, but their assumptions are too strong for most people's liking, so they are next to never used.) I believe fairly strongly that one should not presume knowledge more detailed than the question we are trying to answer. It is much harder (requires larger sample sizes) to estimate variances precisely, so we should not in general imagine that we know whether the variances are equal. I asked you to make inference about means, and if you use the t test that presume equal variances, it is possible that in the presence of unequal sample sizes it might be statistically significant because variances, rather than means, are unequal. I thus use the t test that allows for the possibility of unequal variances.*

After presenting that analysis, I will present the results based on the t test that presumes equal variances.

Methods: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean were tested using a t test that allows for the possibility of unequal variances (Satterthwaite approximation). 95% confidence intervals for the difference in population means were similarly based on that same handling of variances.

Results: Mean serum LDL was 127 mg/dL among the 606 subjects who survived at least 5 years after study enrollment and 119 mg/dL among the 119 subjects who died within 5 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.44 mg/dL to 15.6 mg/dL lower mean LDL among subjects who die within 5 years. Using a t test that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0186), and we can with high confidence reject the null hypothesis that the mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5 years is associated with lower mean serum LDL. *(Note that I get to give a direction in the central tendency for serum LDL levels by vital status. Also, given that I describe the statistical methods previously, I might not again explicitly state that I was using the version of the t test that allowed unequal variances. The t test is so widely used, that I could state the results without reiterating that information. If I were using more complicated statistics (e.g., adjusted regression analyses), I would again remind the reader of the methods.)*

If I use the t test that presumes equal variances, the reporting would differ a little. To be a purist, I will avoid swearing that the t test is testing means. Instead, I will talk about "differences in the distribution"..

Methods: Mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean were tested using a t test that presumes equality of variances. 95% confidence intervals for the difference in population means were similarly based on that same handling of variances.

Results: Mean serum LDL was 127 mg/dL among the 606 subjects who survived at least 5 years after study enrollment and 119 mg/dL among the 119 subjects who died within 5 years. Based on a 95% confidence interval, this observed tendency of 8.50 mg/dL lower mean serum LDL among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.91 mg/dL to 15.1 mg/dL lower mean LDL among subjects who die within 5 years and the variances were identical in the two groups. Using a t test that presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0115), and we can thus conclude with high confidence that the distribution of serum LDL differs between those who do or do not have higher risk of death over a 5 year period. *(Note that I do not get to conclude a direction for the central tendency: The statistical significance could be due to different variances. From the sample descriptive statistics, we see that the group with the smaller sample size (those dying within 5 years) also has greater variability of serum LDL measurements. If that estimated difference in SD of LDL were true in the population, use of the t test that presumes equal variances is anti-conservative: the p values are too low and the CI are too wide. In any case, estimated SD that are higher in the group with lower sample sizes will lead to the p value from the t test that presumes equal sample sizes to be lower than the p value from the t test that allows for the possibility of unequal variances.)*

4. 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.

Instructions for grading: This problem is worth 10 points. Assign points using the same criteria as for problem 3.

Ans: *In choosing how to answer this question, there are basically two options that would typically be considered: the t test that presumes equal variances or the t test that allows for the possibility of unequal variances. In both cases, we would use those t tests on log transformed LDL. The comments made in problem 3 about choosing between these two tests holds here as well.*

Methods: Geometric mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean of log transformed serum LDL levels were tested using a t test that allows for the possibility of unequal variances (Satterthwaite approximation). 95% confidence intervals for the difference in population means for log LDL were similarly based on that same handling of variances. Estimates and CI were then exponentiated in order to obtain inference on the geometric mean.

Results: Geometric mean serum LDL was 123 mg/dL among the 606 subjects who survived at least 5 years after study enrollment and 112 mg/dL among the 119 subjects who died within 5 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 9.65% higher geometric mean among subjects surviving at least 5 years would not be judged unusual if the true ratio of population geometric means indicated anywhere between a 2.01% to 17.9% higher geometric mean LDL among subjects who survive at least 5 years. Using a t test on log transformed LDL that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0128), and we can with high confidence reject the null hypothesis that the geometric mean serum LDL levels are not different by vital status at 5 years in favor of a hypothesis that death within 5

years is associated with lower geometric mean serum LDL. *(Note again my ability to talk about direction of the association in the geometric mean. Also, as before, given that I describe the statistical methods previously, I might not again explicitly state that I was using the version of the t test that allowed unequal variances or that the testing was done on log transformed LDL. There is a reason we put statistical methods in a section that few people read. (cf: Woody on Cheers talking about opera and PBS.))*

If I use the t test that presumes equal variances, the reporting would differ a little. To be a purist, I will avoid swearing that the t test is testing geometric means. Instead, I will talk about "differences in the distribution".

Methods: Geometric mean serum LDL levels were compared between subjects who died within 5 years of study enrollment and those who survived at least 5 years. Differences in the mean of log transformed serum LDL levels were tested using a t test that presumes equal variances. 95% confidence intervals for the difference in population means for log LDL were similarly based on that same handling of variances. Estimates and CI were then exponentiated in order to obtain inference on the geometric mean.

Results: Geometric mean serum LDL was 123 mg/dL among the 606 subjects who survived at least 5 years after study enrollment and 112 mg/dL among the 119 subjects who died within 5 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 9.65% higher geometric mean among subjects surviving at least 5 years would not be judged unusual if the true ratio of population geometric means indicated anywhere between a 3.55% to 16.1% higher geometric mean LDL among subjects who survive at least 5 years and the variances of the log transformed LDL were equal. Using a t test on log transformed LDL that similarly presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P= 0.0016), and we can with high confidence reject the null hypothesis that the distributions of serum LDL levels are not different by vital status at 5 years. *(Because the test of geometric means is ultimately based on the t test, all the comments made in problem 3 apply here as well.)*

5. 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 $\geq$ 160 mg/dL).

Instructions for grading: This problem is worth 10 points. Assign points using the same criteria as for problem 3.

Ans: *In this problem, I am asking for inference about the proportion surviving for 5 years. The choices would be to use differences in proportions or to use ratios of proportions. In either case, the hypothesis test would typically be either the chi squared test or Fisher's exact test.*

(We do have other tests that could be used here, including a likelihood ratio test and a Wald test. The Wald test would be very much like the t test that allows for the possibility of unequal variances. Of course, if the null hypothesis holds, the variances have to be equal, and if an alternative hypothesis holds, the variances have to be unequal. My personal preference in small samples would be to use an unconditional exact test that modifies the Fisther's exact test so it is not so conservative or an unconditional exact test that modifies the chi square test to ensure that it is not anti-conservative. Stata does neither of these, to my knowledge.)

It is far more common to look at differences of proportions, unless the event rate is extremely small. I can use the Stata function cs. (If someone chose to look at ratios of proportions, you can ask me how that would be effected. We will cover it in Lecture 5.) I typically choose the chi squared test over Fisher's exact test.

Methods: **The proportion of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less. Differences in the probability of death within 5 years were tested using Pearson's chi squared test for independence. 95% confidence intervals for the difference in population 5 year mortality probabilities were computed using Wald statistics.**

Results: **Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, 17.0% were observed to die within 5 years, while 13.1% of the subjects with serum LDL greater than or equal to 160 mg/dL died within 5 years of study enrollment. Based on a 95% confidence interval, this 3.91% lower absolute survival probability in subjects with higher serum LDL would not be judged unusual if the true difference in survival probabilities were anywhere between a 10.9% lower absolute probability of survival to a 3.14% higher absolute probability of survival in the high LDL group compared to the low LDL group. Using a chi squared test, this observation is not statistically significant at a 0.05 level of significance (two-sided P= 0.314), and we can not with high confidence reject the null hypothesis that the survival probabilities are not associated with serum LDL levels.** *(If you wanted to quote Fisher's exact test, that two-sided p values was P = 0.396.)*

6. 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 $\geq$ 160 mg/dL).

***Instructions for grading*: This problem is worth 10 points. Assign points using the same criteria as for problem 3.**

Ans: *In this problem, I am asking for inference about the odds of surviving for 5 years, and the odds ratio is the natural comparison across groups. The hypothesis test would typically be either the chi squared test or Fisher's exact test.*

(Again, we have other tests. Because differences in proportions mean that the OR has to be different from 1, in two sample problems, the same tests are used for proportions or odds.)

I will describe an approach based on Fisher's exact test for both the test and the CI. I will report the odds in each group, but that would actually be highly nonstandard. Many people would report the probabilities for each group, while still making inference about the odds ratio.

Methods: **The odds of subjects dying within 5 years of study enrollment were compared between subjects who had serum LDL greater than or equal to 160 mg/dL and subjects whose serum LDL was measured to be 159 mg/dL or less. An odds ratio different from 1 was tested using Fisher's exact test. 95% confidence intervals for the odds ratio was also computed using exact methods.** *(Alternatively, you could have described the use of CI based on Cornfield's method or based on the Wald statistic, which is Woolf's method.)*

Results: **Of the 618 subjects whose serum LDL was less than or equal to 159 mg/dL, the odds of dying within 5 years from study enrollment was 0.205, while for the subjects with serum LDL greater than or equal to 160 mg/dL the odds of 5 year mortality was 0.151. Based on a 95% confidence interval, this observed odds ratio of 0.735 for the comparison of the high LDL group to the low LDL group would not be judged unusual if the true odds ratio were anywhere between 0.373 to 1.36. A Fisher's exact test two-sided p value of 0.396 suggests that we can not with high confidence reject the null hypothesis that the odds of 5 year mortality are not associated with serum LDL levels.** *(The Cornfield CI was 0.406 to 1.33, and the Wald (Woolf) CI was 0.404 to 1.34.)*

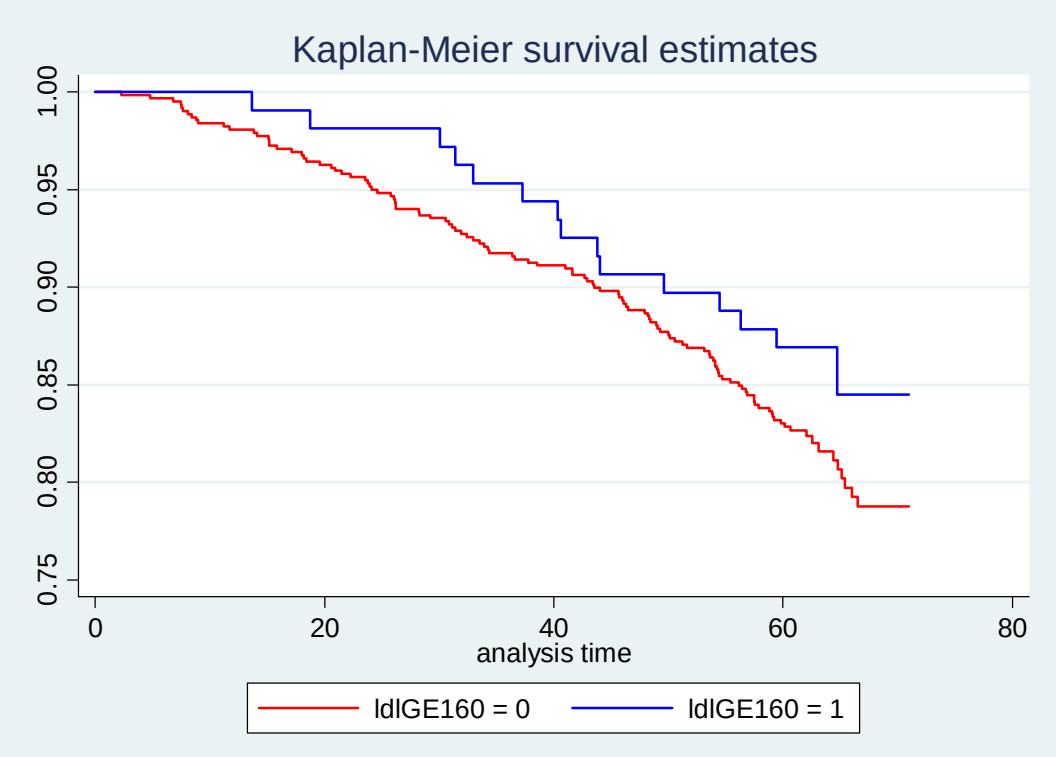
7. Perform a statistical analysis evaluating an association between serum LDL and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum LDL ("high" = LDL $\geq$ 160 mg/dL).

***Instructions for grading*: This problem is worth 10 points. Assign points using much the same criteria as for problem 3.**

Ans: *In this problem, I am asking for inference about the survival distribution as reflected in the hazard function. The typical test would be the logrank test, and estimates would typically be the hazard ratio estimated from proportional hazards regression, because the logrank test corresponds to the score test from proportional hazards regression. (Alternative tests would be the Wilcoxon form of the logrank statistic, and then there is no real estimate of association.)*

Methods: The survival distribution was estimated using Kaplan-Meier estimates with strata defined by serum LDL less than or equal to 159 mg/dL and serum LDL greater than or equal to 160 mg/dL. Difference in survival distributions between those two groups was tested using the logrank statistic. The hazard ratio and 95% CI was computed using Cox proportional hazards regression with the Huber-White sandwich estimator of the standard errors.

Results: The following graph and table depicts Kaplan-Meier estimates of survival probability for the 618 subjects whose serum LDL was less than or equal to 159 mg/dL and the 107 subjects with serum LDL greater than or equal to 160 mg/dL. Apparent from that graph is the tendency for higher survival probailities for the high LDL group at every point in time. The instantaneous risk of death is estimated to be 28.2% lower for the high LDL group compared to the low LDL group. Based on a 95% confidence interval, this observed hazard ratio of 0.718 for the comparison of the high LDL group to the low LDL group would not be judged unusual if the true hazard ratio were anywhere between 0.420 to 1.22. A logrank test two-sided p value of 0.227 suggests that we can not with high confidence reject the null hypothesis that probability of survival is not associated with serum LDL levels.



	Survival Probabilities (Kaplan-Meier)	
	LDL $\leq$ 159 mg/dL	LDL $\geq$ 160 mg/dL
1 years	0.981	1.000
2 years	0.952	0.981
3 years	0.918	0.953
4 years	0.887	0.907
5 years	0.830	0.869

8. 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?

***Instructions for grading*: This problem is worth 10 points. Anyone who invokes choosing an analysis based on the observed P values gets 0 points, no matter what else they write. (But as detailed below, they can talk about the statistical power, which is a tendency to get low P values under the alternative.) Otherwise, assign 2 points for mentioning each of the points I discuss below, and assign up to 4 points for making a final decision consistent with those points. Do not assign a score over 10 points.**

Ans: The correct time to make a decision about which of the above analyses would be used is prior to collecting and analyzing the data. Points that should be considered are:

- **It is scientifically more pleasing to condition on LDL levels and to summarize the survival distribution, if only because the serum LDL measurements must occur earlier in time than the death.**
- **It is statistically much more precise not to have to dichotomize a continuous measurement.**
- ***A priori,* a multiplicative level for LDL levels might be slightly preferred on the basis of biochemistry, but given that the population is not severely diseased, it probably does not make that much of a difference.**
- **The simpler comparisons of means and proportions are probably better understood than the geometric mean, odds ratio, and the hazard ratio (note that the hazard ratio is related to the odds ratio at some technical level).**
- **You have to perform analyses that are valid and that you know how to do.**

All things considered, *a priori* I would have anticipated that of the simple tests, a test of the geometric means across survival groups would have the greatest precision, but that a comparison of means would be nearly as good. Dichotomization of LDL would be expected to perform poorly – even more poorly than dichotomization of survival, because the survival rates are pretty high. By the end of Lecture 6, you will know how to do inference based on treating both survival time and LDL continuously: proportional hazard regression on LDL or (my preference) log LDL.

PROBLEM #1

I check the minimum observation time among the subjects whose time to death was censored. This is about the only reason that I would ever use sample descriptive statistics on a variable that is subject to censoring. The minimum value of 1827 days corresponds to 1827 / 365.25 = 5.002 years.

```
. summ obstime if death==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
     obstime |       602    1945.694    108.4126       1827       2159
```

I thus create a variable to indicate subjects who died within 5 years.

```
. g deadin5= 0
. replace deadin5= 1 if obstime <= 5 * 365.25
(121 real changes made)
```

PROBLEM #2

Descriptive statistics for this problem will consist of the usual descriptive statistics (mean, sd, min, max for continuous random variables and frequencies for binary and categorical random variables) within strata. Two approaches are possible: 1) defining strata based on a categorization of LDL, or 2) defining strata based on the dichotomization of survival at 5 years.

I first create variables that categorize LDL: I divide it into three categories (< 130, 130-160, 160) for the purposes of problem 2, and I divide it into two categories for problems 5-7. I also create a log transformed LDL for use in problem 4.

```
. recode ldl 160/max=3 130/160=2 min/130=1, gen(ldlCTG)
(725 differences between ldl and ldlCTG)

. recode ldl 160/max=1 0/160=0, gen(ldlGE160)
(725 differences between ldl and ldlGE160)

. g logldl= log(ldl)
(10 missing values generated)
```

I also create indicator variables of angina, MI, TIA, CVA, and any CVD.

```
. recode chd 0=0 1=1 2=0, gen(angina)
(91 differences between chd and angina)

. recode chd 0=0 1=0 2=1, gen(mi)
(155 differences between chd and mi)

. recode stroke 0=0 1=1 2=0, gen(tia)
(75 differences between stroke and tia)

. recode stroke 0=0 1=0 2=1, gen(cva)
(99 differences between stroke and cva)

. gen cvd= chd + stroke + chf

. recode cvd 0=0 1/max=1
(cvd: 158 changes made)
```

Stratified statistics within categories of LDL. (I convert these to formal tables in Excel.)

```
. tabstat male age weight packyrs angina mi tia cva chf cvd deadin5,        ///
>         by(ldlCTG) stat(n mean sd min q max) col(stat) long
```

ldlCTG	variable	N	mean	sd	min	p25	p50	p75	max
1	male	393	.5547074	.4976316	0	0	1	1	1
	age	393	74.6972	5.251571	65	71	74	78	92
	weight	393	159.9125	29.93095	86	138.5	160	178	264
	packyrs	393	19.81338	26.94128	0	0	8.4	35	180
	angina	393	.1017812	.3027457	0	0	0	0	1
	mi	393	.1221374	.3278618	0	0	0	0	1
	tia	393	.0381679	.1918458	0	0	0	0	1
	cva	393	.0916031	.2888325	0	0	0	0	1
	chf	393	.0661578	.2488745	0	0	0	0	1
	cvd	393	.3256997	.4692331	0	0	0	1	1
	deadin5	393	.1933842	.395455	0	0	0	0	1
2	male	225	.4311111	.4963358	0	0	0	1	1

	age	225	74.19556	5.624165	67	70	73	77	99
	weight	225	158.384	32.26641	96	137	155	178	245
	packyrs	225	20.00882	28.82649	0	0	6	33.75	240
	angina	225	.0622222	.2420973	0	0	0	0	1
	mi	225	.1244444	.3308239	0	0	0	0	1
	tia	225	.0133333	.1149534	0	0	0	0	1
	cva	225	.1022222	.3036158	0	0	0	0	1
	chf	225	.0488889	.2161165	0	0	0	0	1
	cvd	225	.2577778	.4383863	0	0	0	1	1
	deadin5	225	.1288889	.335824	0	0	0	0	1
3	male	107	.4205607	.4959721	0	0	0	1	1
	age	107	74.8785	5.770305	65	70	74	78	94
	weight	107	162.7402	30.68332	74	143	159	182	257
	packyrs	107	18.08049	24.41289	0	0	3	30	102
	angina	107	.0747664	.2642517	0	0	0	0	1
	mi	107	.1214953	.3282395	0	0	0	0	1
	tia	107	.0560748	.2311487	0	0	0	0	1
	cva	107	.1308411	.3388135	0	0	0	0	1
	chf	107	.0280374	.1658565	0	0	0	0	1
	cvd	107	.317757	.4677955	0	0	0	1	1
	deadin5	107	.1308411	.3388135	0	0	0	0	1
Total	male	725	.4965517	.5003333	0	0	0	1	1
	age	725	74.56828	5.446103	65	71	74	78	99
	weight	725	159.8554	30.77191	74	138.5	158	179	264
	packyrs	725	19.61828	27.16178	0	0	6.5	33.75	240
	angina	725	.0855172	.2798429	0	0	0	0	1
	mi	725	.1227586	.3283865	0	0	0	0	1
	tia	725	.0331034	.1790302	0	0	0	0	1
	cva	725	.1006897	.3011251	0	0	0	0	1
	chf	725	.0551724	.2284741	0	0	0	0	1
	cvd	725	.3034483	.4600645	0	0	0	1	1
	deadin5	725	.1641379	.3706564	0	0	0	0	1

Stratified statistics within categories of mortality. (I convert these to formal tables in Excel.)

```
. tabstat male age weight ldl packyrs angina mi tia cva chf cvd if ldl!=.,   ///
>         by(deadin5) stat(n mean sd min q max) col(stat) long
```

deadin5	vriable	N	mean	sd	min	p25	p50	p75	max
0	male	606	.4669967	.4993217	0	0	0	1	1
	age	606	74.17327	5.20883	65	71	73	77	99
	weight	606	159.9497	30.35261	74	138.5	158.75	179	258
	ldl	606	127.198	32.92893	39	103	127	148	247
	packyrs	606	17.97815	24.71993	0	0	4.35	31.8801	180
	angina	606	.0742574	.262406	0	0	0	0	1
	mi	606	.1006601	.3011266	0	0	0	0	1
	tia	606	.0280528	.1652601	0	0	0	0	1
	cva	606	.0759076	.265069	0	0	0	0	1
	chf	606	.0379538	.1912424	0	0	0	0	1
	cvd	606	.2508251	.4338461	0	0	0	1	1
1	male	119	.6470588	.4799053	0	0	1	1	1
	age	119	76.57983	6.163721	67	72	76	81	91
	weight	119	159.3756	32.95548	96	139	154	177	264
	ldl	119	118.6975	36.15704	11	96	117	142	227
	packyrs	119	27.97056	36.18551	0	0	16.5	46	240
	angina	119	.1428571	.3514067	0	0	0	0	1
	mi	119	.2352941	.4259761	0	0	0	0	1
	tia	119	.0588235	.236289	0	0	0	0	1
	cva	119	.2268908	.4205923	0	0	0	0	1
	chf	119	.1428571	.3514067	0	0	0	0	1
	cvd	119	.5714286	.4969641	0	0	1	1	1
Total	male	725	.4965517	.5003333	0	0	0	1	1
	age	725	74.56828	5.446103	65	71	74	78	99
	weight	725	159.8554	30.77191	74	138.5	158	179	264
	ldl	725	125.8028	33.60197	11	102	125	147	247
	packyrs	725	19.61828	27.16178	0	0	6.5	33.75	240
	angina	725	.0855172	.2798429	0	0	0	0	1
	mi	725	.1227586	.3283865	0	0	0	0	1
	tia	725	.0331034	.1790302	0	0	0	0	1
	cva	725	.1006897	.3011251	0	0	0	0	1
	chf	725	.0551724	.2284741	0	0	0	0	1
	cvd	725	.3034483	.4600645	0	0	0	1	1

PROBLEM #3

T tests comparing mean LDL across groups defined by vital status at 5 years. I personally prefer the t test that allows for the possibility of unequal variances.

```
. ttest ldl, by(deadin5) unequal

Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     606     127.198    1.337646    32.92893     124.571     129.825
       1 |     119    118.6975     3.31451    36.15704    112.1338    125.2611
---------+--------------------------------------------------------------------
combined |     725    125.8028    1.247946    33.60197    123.3527    128.2528
---------+--------------------------------------------------------------------
    diff |            8.500541    3.574252                 1.44132    15.55976
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.3783
Ho: diff = 0                     Satterthwaite's degrees of freedom =  158.746

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9907         Pr(|T| > |t|) = 0.0186          Pr(T > t) = 0.0093
```

I also present the output from a t test that presumes equal variances.

```
. ttest ldl, by(deadin5)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     606     127.198    1.337646    32.92893     124.571     129.825
       1 |     119    118.6975     3.31451    36.15704    112.1338    125.2611
---------+--------------------------------------------------------------------
combined |     725    125.8028    1.247946    33.60197    123.3527    128.2528
---------+--------------------------------------------------------------------
    diff |            8.500541    3.356652                1.910591    15.09049
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.5324
Ho: diff = 0                                     degrees of freedom =      723

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
```

```
 Pr(T < t) = 0.9942         Pr(|T| > |t|) = 0.0115          Pr(T > t) = 0.0058
```

PROBLEM #4

T tests comparing mean log LDL across groups defined by vital status at 5 years. I personally prefer the t test that allows for the possibility of unequal variances.

```
. ttest logldl, by(deadin5) unequal

Two-sample t test with unequal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     606    4.810764    .0109721    .2701018    4.789216    4.832312
       1 |     119    4.718601    .0348673    .3803569    4.649554    4.787648
---------+--------------------------------------------------------------------
combined |     725    4.795636    .0108729    .2927623     4.77429    4.816983
---------+--------------------------------------------------------------------
    diff |            .0921629    .0365529                .0199059    .1644199
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   2.5214
Ho: diff = 0                     Satterthwaite's degrees of freedom =  142.255

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9936         Pr(|T| > |t|) = 0.0128          Pr(T > t) = 0.0064
```

I backtransform the estimates to get geometric means.

```
. di exp(4.810764), exp(4.718601), exp(.0921629), exp(.0199059), exp( .1644199)
122.82542 112.01144 1.0965434 1.0201053 1.1787092
```

I also present the output from a t test that presumes equal variances.

```
. ttest logldl, by(deadin5)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     606    4.810764    .0109721    .2701018    4.789216    4.832312
       1 |     119    4.718601    .0348673    .3803569    4.649554    4.787648
---------+--------------------------------------------------------------------
combined |     725    4.795636    .0108729    .2927623     4.77429    4.816983
```

```
    diff |           .0921629    .0291741                .0348868     .149439
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.1591
Ho: diff = 0                                     degrees of freedom =      723

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9992         Pr(|T| > |t|) = 0.0016          Pr(T > t) = 0.0008
```

Again, I backtransform the estimates to get geometric means.

```
. di exp(4.810764), exp(4.718601), exp(.0921629), exp(.0348868), exp( .149439)
122.82542 112.01144 1.0965434 1.0355025 1.1611826
```

PROBLEM #5

Chi square test comparing proportion of subjects dying within 5 years by LDL greater than or equal to 160. Confidence intervals based on Wald statistics.

```
. cs deadin5 ldlGE160, or

                 | RECODE of ldl          |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |          14         105 |        119
        Noncases |          93         513 |        606
-----------------+------------------------+-----------
           Total |         107         618 |        725
                 |                        |
            Risk |   .1308411    .1699029 |   .1641379
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
-----------------+------------------------+------------------------
 Risk difference |         -.0390618      |   -.1094852    .0313616
      Risk ratio |          .7700935      |    .4585168    1.293396
 Prev. frac. ex. |          .2299065      |   -.2933964    .5414832
 Prev. frac. pop |           .033931      |
      Odds ratio |          .7354839      |    .4068629     1.33063 (Cornfield)
-----------------+------------------------+------------------------
                           chi2(1) =     1.01  Pr>chi2 = 0.3139
```

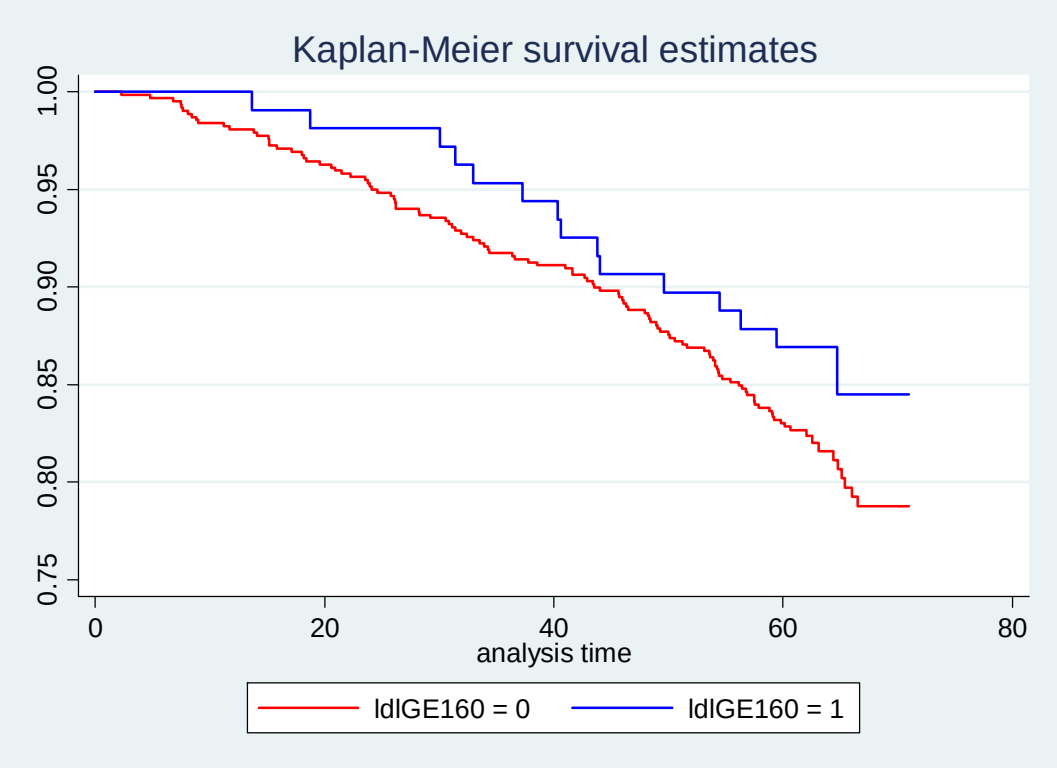
PROBLEM #6

Chi square test comparing odds ratio of subjects dying within 5 years by LDL greater than or equal to 160. Confidence intervals based on Fisher's exact test.

```
. cc deadin5 ldlGE160, exact
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        14         105  |        119      0.1176
        Controls |        93         513  |        606      0.1535
-----------------+------------------------+------------------------
           Total |       107         618  |        725      0.1476
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
-----------------+------------------------+------------------------
      Odds ratio |         .7354839       |    .3726094    1.360582 (exact)
 Prev. frac. ex. |         .2645161       |   -.3605823    .6273906 (exact)
 Prev. frac. pop |         .0405941       |
                 +------------------------+------------------------
                            1-sided Fisher's exact P = 0.1948
                            2-sided Fisher's exact P = 0.3960
```

PROBLEM #7

Descriptive statistics for survival by high vs low LDL. Stratified Kaplan-Meier curves and tabulated survival probabilities.




```
. sts list, by(ldlGE160) at(12 24 36 48 60)

         failure _d:  death
   analysis time _t:  obsmos

                Beg.                    Survivor      Std.
      Time     Total      Fail           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
ldlGE160=0
        12       607        12             0.9806    0.0056     0.9661    0.9889
        24       589        18             0.9515    0.0086     0.9313    0.9658
        36       568        21             0.9175    0.0111     0.8928    0.9366
```

```
      48        549         19              0.8867    0.0127      0.8590    0.9093
      60        514         35              0.8301    0.0151      0.7981    0.8575
ldlGE160=1
      12          0          0              1.0000         .           .         .
      24        106          2              0.9813    0.0131      0.9273    0.9953
      36        103          3              0.9533    0.0204      0.8914    0.9803
      48         98          5              0.9065    0.0281      0.8333    0.9486
      60         94          4              0.8692    0.0326      0.7891    0.9203
```

Logrank test comparing risk of death by high vs low LDL along with hazard ratio estimate from Cox proportional hazards regression model and Wald based CI.

```
. sts test ldlGE160

         failure _d:  death
   analysis time _t:  obsmos


Log-rank test for equality of survivor functions

         |   Events         Events
ldlGE160 |  observed       expected
---------+------------------------
0        |       116         111.01
1        |        15          19.99
---------+------------------------
Total    |       131         131.00

               chi2(1) =       1.47
               Pr>chi2 =     0.2249


. stcox ldlGE160, robust

         failure _d:  death
   analysis time _t:  obsmos

Cox regression -- Breslow method for ties

No. of subjects      =          725                Number of obs   =       725
```

```
No. of failures      =          131
Time at risk         =  43077.76313
                                                Wald chi2(1)    =      1.47
Log pseudolikelihood =   -839.53753             Prob > chi2     =    0.2256
------------------------------------------------------------------------------
             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ldlGE160 |   .7178716   .1963807    -1.21   0.226     .4199449     1.22716
------------------------------------------------------------------------------


. stcox ldlGE160

        failure _d:  death
  analysis time _t:  obsmos

Cox regression -- Breslow method for ties

No. of subjects =          725                  Number of obs   =       725
No. of failures =          131
Time at risk    =  43077.76313
                                                LR chi2(1)      =      1.59
Log likelihood  =   -839.53753                  Prob > chi2     =    0.2076
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ldlGE160 |   .7178716   .1969881    -1.21   0.227     .4192492    1.229197
------------------------------------------------------------------------------
```