Biost 518: Applied Biostatistics II
Biost 515: Biostatistics II
Emerson, Winter 2015

Homework #1 Key
January 5, 2015

***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 dropbox by 9:30 am on Monday, January 12, 2015. 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 #1 from 2014 or HWs #1, 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 was new last year, and thus keys prior to 2014 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 C reactive protein (CRP) 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 inflammatory biomarkers and mortality. The data can be found on the class web page (follow the link to Datasets) in the file labeled inflamm.txt. Documentation is in the file inflamm.pdf. The data is in free-field format, and can be read into R by

```
read.table("http://www.emersonstatistics.com/datasets/inflamm.txt",header=T)
```

It can be read into Stata using the following code in a .do file.

```
infile id site age male bkrace smoker estrogen prevdis diab2 bmi ///
      systBP aai cholest crp fib ttodth death cvddth                    ///
      using http://www.emersonstatistics.com/datasets/inflamm.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 CRP levels are as follows (taken from the Mayo Clinic website):

Below 1 mg/L	Low risk of heart disease
1 - 3 mg/L	Average risk of heart disease
Above 3 mg/L	High risk of heart disease

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 4 years of study enrolment or death after 4 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 1,480 days, or just over 4 years. Hence the vital status of every individual is known at 4 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 CRP and 4 year all-cause mortality in the medical literature. In addition to the two variables of primary interest, you may restrict attention to age, sex, BMI, smoking history, cholesterol, and prior history of cardiovascular disease.

***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. (CRP 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 CRP). 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), and 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 columns (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: An indicator variable was created for death within 4 years of study enrollment (no subjects were censored during that period of observation). Data from 67 subjects missing data for serum C reactive protein (CRP) was excluded from the analysis. Subjects missing data for any other variable were excluded only from the analyses involving those variables. Descriptive statistics are presented within groups defined by serum CRP measurements (less than 1 mg/L, between 1 and 3 mg/L inclusive, and greater than 3 mg/L), as well as in the entire sample that had CRP measurements available. Within each group defined by serum CRP level, for continuous variables (age, body mass index (BMI), cholesterol) we include the mean, standard deviation, minimum and maximum. For binary variables (sex and indicators of smoking, prior history of cardiovascular disease (CVD), or death) we present percentages.

Results: Data is available on 5,000 subjects, however 67 of those subjects (including 11 who died within 4 years) are missing data on serum C reactive protein (CRP). 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. In addition, of the subjects with available data for CRP, 6 subjects were missing data on smoking status, 13 subjects were missing data on BMI, and 3 subjects were missing data on serum cholesterol. None of these subjects had missing data for more than one variable.

Of the 4,933 subjects with available CRP measurements, 428 had serum CRP measurements less than 1 mg/L (all recorded as 0 mg/L), 3,330 had measurements between 1 and 3 mg/L inclusive, and 1,175 had measurements greater than 3 mg/L. The following table (Table 2a) presents descriptive statistics within these groups. Several trends were evident across the categories defined by serum CRP levels: With higher serum CRP, the subjects were less likely to be male, more likely to have higher BMI, more likely to smoke, and more likely to have had prior CVD. There were less evidence of trends seen across groups in age or cholesterol. Subjects with the lowest levels of serum CRP appeared to have a lower mortality rate: 4.9% of subjects with CRP less than 1 mg/L died within 4 years compared to about 8.4% in the middle group and 15.6% in the group having highest serum CRP at study entry.

Table 2a: Descriptive statistics for selected variables in subjects with available CRP measurements.

	Serum C Reactive Protein (CRP)			
	< 1 mg/L (n=428) [2]	**1 - 3 mg/L (n=3,330) [2]**	**> 3 mg/dL (n=1,175) [2]**	**Any Level (n=4,933)**
Male (%)	45.6%	43.3%	37.0%	42.0%
Age (yrs) [1]	73.5 (5.80; 65 - 94)	72.7 (5.52; 65 - 100)	72.7 (5.58; 65 - 93)	72.8 (5.56; 65 - 100)
Body mass index (BMI) (kg/m^2) [1]	23.8 (3.64; 15.6 - 38.6)	26.4 (4.31; 14.7 - 53.2)	28.5 (5.46; 15.3 - 58.8)	26.7 (4.72; 14.7 - 58.8)
Smoker (%)	9.6%	11.0%	16.4%	12.2%
Serum cholesterol (mg/dL) [1]	206 (40.5; 109 - 407)	213 (38.6; 73 - 363)	211 (40.4; 97 - 430)	212 (39.2; 73 - 430)

Prior cardiovascular disease (%)	18.2%	21.5%	28.8%	22.9%
Death w/in 4 years	4.9%	8.4%	15.6%	9.8%

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

[2] There were 67 subjects missing data for CRP, and no data for these subjects were included in the table. In addition, some subjects were missing data for other variables. In the low CRP group, 1 subject was missing data for smoking status and 1 subject was missing data for cholesterol. In the middle CRP group, 12 subjects were missing data for BMI and 5 subjects were missing data for smoking status. In the high CRP group, 1 subject was missing data for BMI, and 2 subjects were missing data for cholesterol.

Below I also provide an answer based on dividing the sample according to survival for at least 4 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 CRP 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: An indicator variable was created for death within 4 years of study enrollment (no subjects were censored during that period of observation). Data from 67 subjects missing data for serum C reactive protein (CRP) was excluded from the analysis. Subjects missing data for any other variable were excluded only from the analyses involving those variables. Descriptive statistics are presented within groups defined by death within 4 years, survival for 4 years post study entry, and for the entire sample having available CRP measurements. Within each group defined by vital status at 4 years, for continuous variables (age, body mass index (BMI), cholesterol, serum CRP) we include the mean, standard deviation, minimum and maximum. For binary variables (sex and indicators of smoking, prior history of cardiovascular disease (CVD), or death) we present percentages.

Results: Data is available on 5,000 subjects, however 67 of those subjects (including 11 who died within 4 years) are missing data on serum C reactive protein (CRP). 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. In addition, of the subjects with available data for CRP, 6 subjects were missing data on smoking status, 13 subjects were missing data on BMI, and 3 subjects were missing data on serum cholesterol. None of these subjects had missing data for more than one variable.

Of the 4,933 subjects with available CRP measurements, 484 (9.8%) died within 4 years of study enrollment and 4,449 were still alive 4 years after study enrollment. The following table (Table 2b) presents descriptive statistics within these groups. Subjects dying within 4 years were more likely to be male, tended to be older, were more likely to be smokers, tended to have lower serum cholesterol, and were more likely to have prior history of cardiovascular disease than subjects surviving for at least 4 years after study enrollment. The average BMI was similar across the groups defined by vital status after 4 years. Subjects dying within 4 years also tended toward higher serum CRP at study enrollment: mean serum CRP was 5.38 mg/L in those observed to die within 4 years compared to a mean serum CRP of 3.42 mg/L in those surviving at least 4 years.

Table 2b: Descriptive statistics for selected variables in subjects with available CRP measurements.

	Vital Status at 4 Years Post Study Enrollment		
	Alive at 4 Years (n=4,449)	**Death w/in 4 Years (n=484)**	**All Subjects (n=4,933)**
Male (%)	40.0%	60.1%	42.0%
Age (yrs) [1]	72.4 (5.29; 65 - 98)	76.2 (6.70; 65 - 100)	72.8 (5.56; 65 - 100)
Body mass index (BMI) (kg/m^2) [1]	26.7 (4.69; 14.7 - 58.8)	26.3 (4.98; 14.8 - 48.1)	26.7 (4.72; 14.7 - 58.8)
Smoker (%)	12.0%	14.3%	12.2%

Serum cholesterol (mg/dL) [1]	213 (38.9; 78 - 430)	204 (41.4; 73 - 396)	212 (39.2; 73 - 430)
Serum C reactive protein (mg/L) [1]	3.42 (5.87; 0 - 108)	5.38 (8.10; 0 - 55)	3.61 (6.15; 0 - 108)
Prior cardiovascular disease (%)	20.9%	41.9%	22.9%

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

[2] There were 67 subjects missing data for CRP, and no data for these subjects were included in the table. In addition, some subjects were missing data for other variables. In the subjects still alive 4 years after study enrolment, 11 subjects were missing data for BMI, 6 subjects were missing data for smoking status, and 3 subjects were missing data for cholesterol. In thesubjects who died within 4 years, 2 subjects were missing data for BMI.

3. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing mean CRP values across groups defined by vital status at 4 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 CRP levels were compared between subjects who died within 4 years of study enrollment and those who survived at least 4 years. Differences in the mean were tested using a t test that allows for the possibility of unequal variances (Satterthwaite approximation), with two-sided p values. 95% confidence intervals for the difference in population means were similarly based on that same handling of variances. Subjects missing data for CRP were excluded from the analysis.

Results: Mean serum CRP was 3.42 mg/L among the 4,449 subjects who survived at least 4 years after study enrollment and 5.38 mg/L among the 484 subjects who died within 4 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 1.95 mg/dL higher mean serum CRP among subjects dying earlier

would not be judged unusual if the true difference population means were anywhere between a 1.21 mg/L to 2.70 mg/L higher mean CRP among subjects who die within 4 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< 0.0001), and we can with high confidence reject the null hypothesis that the mean serum CRP levels are not different by vital status at 4 years in favor of a hypothesis that death within 4 years is associated with higher mean serum CRP. *(Note that I get to give a direction in the central tendency for serum CRP 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 CRP levels were compared between subjects who died within 4 years of study enrollment and those who survived at least 4 years. Differences in the mean were tested using a t test that presumes equality of variances, with two-sided p values. 95% confidence intervals for the difference in population means were similarly based on that same handling of variances. Subjects missing data for CRP were excluded from the analysis.

Results: Mean serum CRP was 3.42 mg/L among the 4,449 subjects who survived at least 4 years after study enrollment and 5.38 mg/L among the 484 subjects who died within 4 years. Based on a 95% confidence interval, this observed tendency of 1.95 mg/L higher mean serum CRP among subjects dying earlier would not be judged unusual if the true difference population means were anywhere between a 1.38 mg/L to 2.53 mg/L higher mean CRP among subjects who die within 4 years providing 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< .0001), and we can thus conclude with high confidence that the distribution of serum CRP differs between those who do or do not have higher risk of death over a 4 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 4 years) also has much greater variability of serum CRP measurements. If that estimated difference in SD of CRP 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 narrow. 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, and the CI when assuming equal variances will be narrower than the CI when allowing for unequal variances.)*

4. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing geometric mean CRP values across groups defined by vital status at 4 years. (Note that there are some measurements of CRP that are reported as zeroes. Make clear how you handle these measurements.)

Instructions for grading: This problem is worth 10 points. Assign points using the same criteria as for problem 3, with the additional requirement that the student explicitly state how he/she handled the measurements of CRP reported as 0.

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 CRP. The comments made in problem 3 about choosing between these two tests holds here as well.

With regard to using geometric means when some subjects have measurements recorded as 0, we must consider how to handle these. Various approaches have been used by different people:

- *When we believe the zero values are merely representative of a positive measurement below some limit of detection, we might consider "imputing" what those values might truly have tended to be. A simple (simplistic?) approach in this setting might be to assign all such measurements to be equal to one-half the lower limit of detection. When we do not know exactly what that limit might be, we could consider one-half the lowest observed value. The key point is that we do not want to go to some extreme like using 0.00000001, because that will produce quite influential outliers after taking the log. Note also that this "single imputation" approach does not model the variability that might have been present among those observations, but it is not really any worse than "round-off error" in the recording of the other measurements (that is, the subjects having recorded CRP of 1 mg/L were likely truly many different measurements*
- *Some people might separately analyze the probability to have a zero measurement, and then the geometric mean of the nonzero measurements. This could be especially attractive when you know that the zeroes are true zeroes. (For instance, for many people, the number of cigarettes smoked is truly zero, not just something below a detectable limit.) But this method has the drawback of having to test two different parameters to detect a difference in distributions.*
- *Some people try shifting all measurements by adding 1 to <u>every</u> measurement. Personally I do not think that this is as reasonable as the prior approaches.*

<u>Methods:</u> Geometric mean serum CRP levels were compared between subjects who died within 4 years of study enrollment and those who survived at least 4 years. Differences in the mean of log transformed serum CRP 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 CRP were similarly based on that same handling of variances. Estimates and CI were then exponentiated in order to obtain inference on the geometric mean. Subjects with recorded CRP of 0 mg/L were presumed to have some positive value below a lower limit of detection, and a value of 0.5 mg/L was imputed for all such subjects. Subjects missing data for CRP were excluded from the analysis.

<u>Results:</u> Geometric mean serum CRP was 2.03 mg/L among the 4,449 subjects who survived at least 4 years after study enrollment and 2.97 mg/L among the 484 subjects who died within 4 years. Based on a 95% confidence interval computed with an allowance for unequal variances, this observed tendency of 46.4% higher geometric mean among subjects dying within 4 years would not be judged unusual if the true ratio of population geometric means indicated anywhere between a 33.2% to 60.9% higher geometric mean CRP among subjects dying within 4 years. Using a t test on log transformed CRP that similarly allows for the possibility of unequal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P< .0001), and we can with high confidence reject the null hypothesis that the geometric mean serum CRP levels are not different by vital status at 4 years in favor of a hypothesis that death within 4 years is associated with higher geometric mean serum CRP. *(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 CRP. 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 CRP levels were compared between subjects who died within 4 years of study enrollment and those who survived at least 4 years. Differences in the mean of log transformed serum CRP levels were tested using a t test that presumes equal variances. 95% confidence intervals for the difference in population means for log CRP were similarly based on that same handling of variances. Estimates and CI were then exponentiated in order to obtain inference on the geometric mean. Subjects with recorded CRP of 0 mg/L were presumed to have some positive value below a lower limit of detection, and a value of 0.5 mg/L was imputed for all such subjects. Subjects missing data for CRP were excluded from the analysis.

Results: Geometric mean serum CRP was 2.03 mg/L among the 4,449 subjects who survived at least 4 years after study enrollment and 2.97 mg/L among the 484 subjects who died within 4 years. Based on a 95% confidence interval computed by presuming equal variances, this observed tendency of 46.4% higher geometric mean among subjects dying within 4 years would not be judged unusual if the true ratio of population geometric means indicated anywhere between a 34.1% to 59.8% higher geometric mean CRP among subjects who die within 4 years and the variances of the log transformed CRP were equal. Using a t test on log transformed CRP that similarly presumes equal variances, this observation is statistically significant at a 0.05 level of significance (two-sided P< 0.0001), and we can with high confidence reject the null hypothesis that the distributions of serum CRP levels are not different by vital status at 4 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 CRP and 4 year all-cause mortality by comparing the probability of death within 4 years across groups defined by whether the subjects have high serum CRP ("high" = CRP > 3 mg/L).

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 4 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 4 years of study enrollment were compared between subjects who had serum CRP greater than 3 mg/L and subjects whose serum CRP was measured to be 3 mg/L or less. Differences in the probability of death within 4 years were tested using Pearson's chi squared test for independence. 95% confidence

intervals for the difference in population 4 year mortality probabilities were computed using Wald statistics. Subjects missing data for CRP were excluded from the analysis.

Results: Of the 3,758 subjects whose serum CRP was less than or equal to 3 mg/L, 8.0% were observed to die within 4 years, while 15.6% of the 1,175 subjects with serum CRP greater than 3 mg/L died within 4 years of study enrollment. Based on a 95% confidence interval, this 7.56% higher absolute mortality in subjects with higher serum CRP would not be judged unusual if the true difference in mortalities were anywhere between a 5.32% higher absolute mortality to a 9.81% higher absolute mortality in the high CRP group compared to the low CRP group. Using a chi squared test, this observation is not statistically significant at a 0.05 level of significance (two-sided P< .0001), and we can with high confidence reject the null hypothesis that the survival probabilities are not associated with serum CRP levels. *(If you wanted to quote Fisher's exact test, that two-sided p values was also P < .0001.)*

6. Perform a statistical analysis evaluating an association between serum CRP and 4 year all-cause mortality by comparing the odds of death within 4 years across groups defined by whether the subjects have high serum CRP ("high" = CRP > 3 mg/L).

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 4 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 4 years of study enrollment were compared between subjects who had serum CRP greater than 3 mg/L and subjects whose serum CRP was measured to be 3 mg/L 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. Subjects missing data for CRP were excluded from the analysis. *(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 3,758 subjects whose serum CRP was less than or equal to 3 mg/L, the odds of dying within 4 years from study enrollment was 0.0871, while for the subjects with serum CRP greater 3 mg/L the odds of 4 year mortality was 0.184. Based on a 95% confidence interval, this observed odds ratio of 2.12 for the comparison of the high CRP group to the low CRP group would not be judged unusual if the true odds ratio were anywhere between 1.73 to 2.59. A Fisher's exact test two-sided p value of P < .0001 suggests that we can with high confidence reject the null hypothesis that the odds of 4 year mortality is associated with serum CRP levels. *(The Cornfield CI was 1.74 to 2.58, and the Wald (Woolf) CI was 1.75 to 2.58. In these large sample sizes, there is little difference between the results from the alternative methods for CI)*

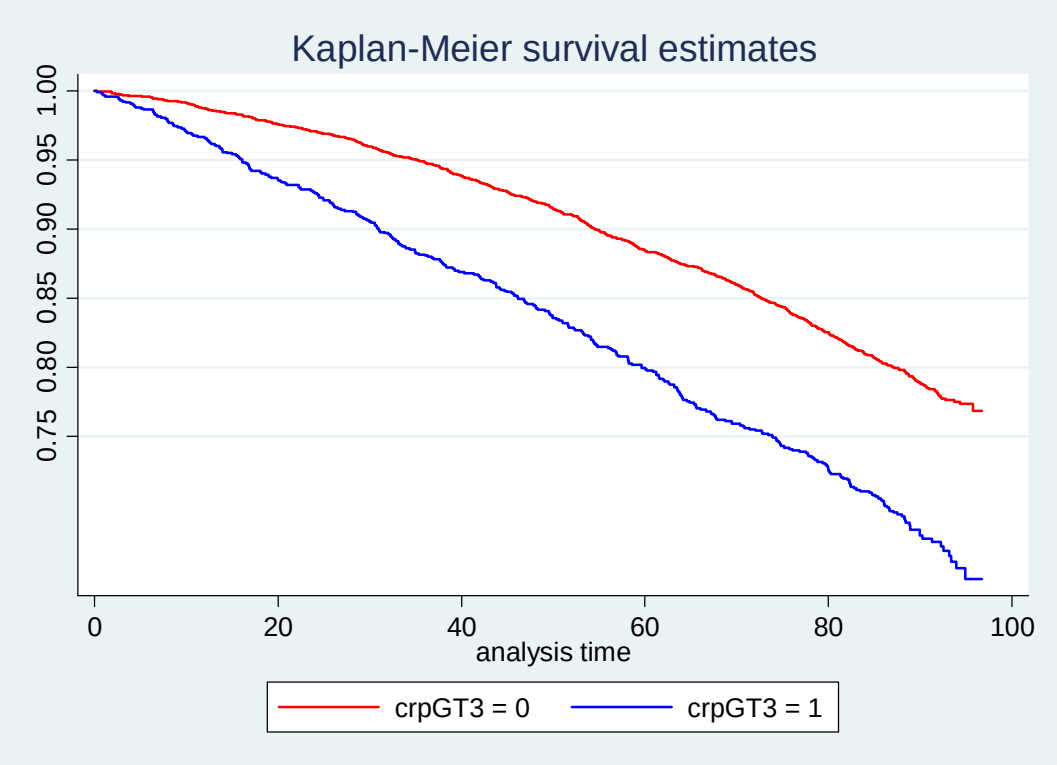
7. Perform a statistical analysis evaluating an association between serum CRP and all-cause mortality over the entire period of observation of these subjects by comparing the instantaneous risk of death across groups defined by whether the subjects have high serum CRP ("high" = CRP > 3 mg/L).

***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.) Note that I realize that you did not learn to use the HR estimate in this way last quarter (I think this is an undesirable omission), but you certainly were taught to use the Kaplan-Meier estimates.*

Methods: The survival distribution was estimated using Kaplan-Meier estimates with strata defined by serum CRP less than or equal to 3 mg/L and serum CRP greater than 3 mg/L. 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. Subjects missing data for CRP were excluded from the analysis.

Results: The following graph and table depicts Kaplan-Meier estimates of survival probability for the 3,758 subjects whose serum CRP was less than or equal to 3 mg/L and the 1,175 subjects with serum CRP greater than 3 mg/L. Apparent from that graph is the tendency for higher survival probailities for the low CRP group at every point in time. The instantaneous risk of death is estimated to be 68.7% higher for the high CRP group compared to the low CRP group. Based on a 95% confidence interval, this observed hazard ratio of 1.687 for the comparison of the high CRP group to the low CRP group would not be judged unusual if the true hazard ratio were anywhere between 1.486 and 1.917. A logrank test two-sided p value of P < 0.0001 suggests that we can with high confidence reject the null hypothesis that probability of survival is not associated with serum CRP levels.



	Survival Probabilities (Kaplan-Meier)	
	CRP $\leq$ 3 mg/L	CRP > 3 mg/L
1 years	0.988	0.967
2 years	0.971	0.926
3 years	0.948	0.881
4 years	0.920	0.844
5 years	0.884	0.800

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 CRP? 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 CRP levels and to summarize the survival distribution, if only because the serum CRP 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 CRP levels might be greatly preferred on the basis of biochemistry, and the fact that CRP is at very low levels in the absence of inflammation. In the presence of inflammation, it probably behaves multiplicatively.**
- **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 CRP 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 CRP continuously: proportional hazard regression on CRP or (my preference) log CRP.

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 1480 days corresponds to 1480 / 365.25 = 4.052 years.

```
. summ ttodth if death==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      ttodth |      3879    2603.711    413.5922       1480       2942
```

I thus create a variable to indicate subjects who died within 4 years.

```
. g deadin4= 0
. replace deadin4= 1 if ttodth <= 4 * 365.25
(495 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 CRP, or 2) defining strata based on the dichotomization of survival at 4 years.

I first create variables that categorize CRP: I divide it into three categories for the purposes of problem 2, and I divide it into two categories for problems 5-7. I also create a log transformed CRP for use in problem 4. Note the handling of the values with CRP==0.

```
. recode crp 0=0 1/3=1 4/max=2, gen(crpCTG)
(2964 differences between crp and crpCTG)

. recode crp 4/max=1 0/3=0, gen(crpGT3)
(4505 differences between crp and crpGT3)

. g logcrp= log(crp)
(495 missing values generated)

. replace logcrp= log(0.5) if crp==0
(428 real changes made)
```

I investigate missing data patterns.

```
. table deadin4 if crp==.

  deadin4 |      Freq.
----------+-----------
        0 |         56
        1 |         11
```

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

```
. tabstat male age bmi smoker cholest prevdis deadin4,         ///
>         by(crpCTG) stat(n mean sd min q max) col(stat) long
```

crpCTG	variable	N	mean	sd	min	p25	p50	p75	max
0	male	428	.4556075	.4986082	0	0	0	1	1
	age	428	73.45093	5.80357	65	69	72	77	94
	bmi	428	23.81636	3.639352	15.6	21.25	23.4	25.9	38.6
	smoker	427	.0960187	.2949625	0	0	0	0	1
	cholest	427	205.9953	40.52681	109	181	204	231	407
	prevdis	428	.182243	.3864964	0	0	0	0	1
	deadin4	428	.0490654	.2162574	0	0	0	0	1
1	male	3330	.433033	.4955695	0	0	0	1	1
	age	3330	72.73694	5.523851	65	68	72	76	100
	bmi	3318	26.38927	4.306228	14.7	23.5	26	28.7	53.2
	smoker	3325	.1100752	.3130305	0	0	0	0	1
	cholest	3330	212.8279	38.57278	73	187	211	237	363
	prevdis	3330	.2147147	.410686	0	0	0	0	1
	deadin4	3330	.0840841	.2775556	0	0	0	0	1
2	male	1175	.3702128	.4830672	0	0	0	1	1
	age	1175	72.73532	5.58111	65	68	71	76	93
	bmi	1174	28.45494	5.463167	15.3	24.5	27.6	31.5	58.8
	smoker	1175	.1642553	.3706649	0	0	0	0	1
	cholest	1173	210.5021	40.38732	97	184	209	236	430
	prevdis	1175	.2876596	.4528643	0	0	0	1	1
	deadin4	1175	.1557447	.3627675	0	0	0	0	1
Total	male	4933	.4200284	.4936131	0	0	0	1	1
	age	4933	72.7985	5.564772	65	68	72	76	100
	bmi	4920	26.65835	4.721579	14.7	23.5	26.1	29.15	58.8
	smoker	4927	.121778	.3270624	0	0	0	0	1
	cholest	4930	211.6828	39.22629	73	186	210	236	430
	prevdis	4933	.2292722	.4204073	0	0	0	0	1

```
             deadin4 |      4933   .0981147   .2974999          0          0          0          0          1

. list bmi smoker cholest if crp!=. & crp==0 & (bmi==. | smoker==. | cholest==.)

     +-------------------------+
     |  bmi   smoker   cholest |
     |-------------------------|
 404. | 26.1        0         . |
4548. | 21.1        .       142 |
     +-------------------------+

. list bmi smoker cholest if crp!=. & crp>=1 & crp<=3 & (bmi==. | smoker==. | cholest==.)

     +-------------------------+
     |  bmi   smoker   cholest |
     |-------------------------|
  52. |    .        0       244 |
 102. |    .        0       195 |
 267. | 29.2        .       207 |
 714. |    .        0       192 |
 996. |    .        0       156 |
     |-------------------------|
1000. |    .        0       259 |
1409. | 28.7        .       205 |
1495. |    .        0       199 |
2863. | 25.2        .       188 |
3324. | 24.5        .       192 |
     |-------------------------|
3424. |    .        0       193 |
3840. |    .        0       226 |
4007. |    .        0       246 |
4143. |    .        0       264 |
4286. | 37.8        .       190 |
     |-------------------------|
4902. |    .        0       194 |
4991. |    .        0       282 |
     +-------------------------+

. list bmi smoker cholest if crp!=. & crp>3 & (bmi==. | smoker==. | cholest==.)

     +-------------------------+
```

```
      |  bmi   smoker   cholest |
      |-------------------------|
 803. |    .        1       188 |
3564. | 26.2        0         . |
3566. | 31.1        1         . |
      +-------------------------+
```

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

```
.
. tabstat male age bmi smoker cholest crp prevdis if crp!=.,   ///
>         by(deadin4) stat(n mean sd min q max) col(stat) long
```

deadin4	variable	N	mean	sd	min	p25	p50	p75	max
0	male	4449	.4003147	.4900172	0	0	0	1	1
	age	4449	72.42392	5.294232	65	68	71	76	98
	bmi	4438	26.69498	4.691815	14.7	23.5	26.1	29.2	58.8
	smoker	4443	.1195138	.3244287	0	0	0	0	1
	cholest	4446	212.5045	38.89582	78	187	211	236	430
	crp	4449	3.422117	5.871972	0	1	2	3	108
	prevdis	4449	.2085862	.4063436	0	0	0	0	1
1	male	484	.6012397	.4901499	0	0	1	1	1
	age	484	76.24174	6.701676	65	71	76	81	100
	bmi	482	26.32116	4.979752	14.8	23.2	25.6	28.8	48.1
	smoker	484	.142562	.3499874	0	0	0	0	1
	cholest	484	204.1343	41.42664	73	176	201.5	229	396
	crp	484	5.376033	8.097691	0	1	3	6	55
	prevdis	484	.4194215	.493975	0	0	0	1	1
Total	male	4933	.4200284	.4936131	0	0	0	1	1
	age	4933	72.7985	5.564772	65	68	72	76	100
	bmi	4920	26.65835	4.721579	14.7	23.5	26.1	29.15	58.8
	smoker	4927	.121778	.3270624	0	0	0	0	1
	cholest	4930	211.6828	39.22629	73	186	210	236	430
	crp	4933	3.613825	6.152715	0	1	2	3	108
	prevdis	4933	.2292722	.4204073	0	0	0	0	1

PROBLEM #3

T tests comparing mean CRP across groups defined by vital status at 4 years. I personally prefer the t test that allows for the possibility of unequal variances.

```
. ttest crp, by(deadin4) unequal

Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4449    3.422117    .0880345    5.871972    3.249526    3.594709
       1 |     484    5.376033    .3680768    8.097691    4.652803    6.099263
---------+--------------------------------------------------------------------
combined |    4933    3.613825    .0876014    6.152715    3.442087    3.785563
---------+--------------------------------------------------------------------
    diff |           -1.953916    .3784582               -2.697348   -1.210484
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -5.1628
Ho: diff = 0                     Satterthwaite's degrees of freedom =  539.648

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

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

```
. ttest crp, by(deadin4)

Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4449    3.422117    .0880345    5.871972    3.249526    3.594709
       1 |     484    5.376033    .3680768    8.097691    4.652803    6.099263
---------+--------------------------------------------------------------------
combined |    4933    3.613825    .0876014    6.152715    3.442087    3.785563
---------+--------------------------------------------------------------------
    diff |           -1.953916    .2932011               -2.52872   -1.379111
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -6.6641
Ho: diff = 0                                     degrees of freedom =     4931

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

PROBLEM #4

T tests comparing mean log CRP across groups defined by vital status at 4 years. I personally prefer the t test that allows for the possibility of unequal variances.

```
. ttest logcrp, by(deadin4) unequal

Two-sample t test with unequal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
```

```
       0 |    4449    .7078204    .0139076    .9276468    .6805546    .7350862
       1 |     484    1.088892    .0461802    1.015964    .9981534    1.179631
combined |    4933    .7452092    .0134321    .9434071    .7188763     .771542
    diff |           -.3810718    .0482289               -.4757985   -.2863452
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -7.9013
Ho: diff = 0                     Satterthwaite's degrees of freedom =  574.073

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

I backtransform the estimates to get geometric means.

```
. di exp(.7078204), exp(1.088892), exp(.3810718), exp(.4757985), exp( .2863452)
2.0295628 2.9709804 1.4638527 1.6092987 1.331552
```

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

```
. ttest logcrp, by(deadin4)

Two-sample t test with equal variances
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    4449    .7078204    .0139076    .9276468    .6805546    .7350862
       1 |     484    1.088892    .0461802    1.015964    .9981534    1.179631
---------+--------------------------------------------------------------------
combined |    4933    .7452092    .0134321    .9434071    .7188763     .771542
---------+--------------------------------------------------------------------
    diff |           -.3810718    .0448318               -.4689621   -.2931815
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -8.5000
Ho: diff = 0                                     degrees of freedom =     4931

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Again, I backtransform the estimates to get geometric means.

```
. di exp(.7078204), exp(1.088892), exp(.3810718), exp(.4689621), exp(.2931815)
2.0295628 2.9709804 1.4638527 1.5983344 1.3406861
```

PROBLEM #5

Chi square test comparing proportion of subjects dying within 4 years by CRP greater than or equal to 160. Confidence intervals based on Wald statistics.

```
. cs deadin4 crpGT3, or

                | RECODE of crp          |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
          Cases |       183         301  |        484
       Noncases |       992        3457  |       4449
----------------+------------------------+-----------
          Total |      1175        3758  |       4933
                |                        |
           Risk |  .1557447    .0800958  |   .0981147
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
----------------+------------------------+------------------------
Risk difference |         .0756489       |    .0531723    .0981254
     Risk ratio |          1.94448       |    1.637791    2.308599
 Attr. frac. ex. |         .4857237       |    .3894216    .5668368
 Attr. frac. pop |         .1836517       |
     Odds ratio |          2.118714      |    1.740572    2.579029 (Cornfield)
----------------+------------------------+------------------------
                            chi2(1) =    57.89  Pr>chi2 = 0.0000
```

PROBLEM #6

Chi square test comparing odds ratio of subjects dying within 4 years by CRP greater than or equal to 160. Confidence intervals based on Fisher's exact test.

```
. cc deadin4 crpGT3, exact
                                                      Proportion
                |   Exposed   Unexposed  |      Total     Exposed
----------------+------------------------+-----------------------
          Cases |       183         301  |        484      0.3781
       Controls |       992        3457  |       4449      0.2230
----------------+------------------------+-----------------------
          Total |      1175        3758  |       4933      0.2382
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
----------------+------------------------+------------------------
     Odds ratio |          2.118714      |    1.729676    2.591835 (exact)
 Attr. frac. ex. |         .5280155       |     .421857     .614173 (exact)
 Attr. frac. pop |         .1996422       |
                +-------------------------------------------------
                         1-sided Fisher's exact P = 0.0000
                         2-sided Fisher's exact P = 0.0000
```

PROBLEM #7

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

```
. g obsmos= ttodth / 30.4

. stset obsmos death

     failure event:  death != 0 & death < .
obs. time interval:  (0, obsmos]
 exit on or before:  failure

--------------------------------------------------------------------------------
     5000  total obs.
        0  exclusions
--------------------------------------------------------------------------------
     5000  obs. remaining, representing
     1121  failures in single record/single failure data
 389536.4  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =  96.77631

. sts graph, by(crpGT3) ylabel(0.75(.05)1.00)       ///
>         plot1opts(lcolor(red)) plot2opts(lcolor(blue))

         failure _d:  death
   analysis time _t:  obsmos

. sts list, by(crpGT3) at(12 24 36 48 60)

         failure _d:  death
   analysis time _t:  obsmos

              Beg.                    Survivor      Std.
    Time     Total     Fail            Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
crpGT3=0
      12      3712       47             0.9875    0.0018     0.9834    0.9906
      24      3650       63             0.9707    0.0027     0.9648    0.9757
      36      3565       84             0.9484    0.0036     0.9408    0.9550
      48      3458      107             0.9199    0.0044     0.9108    0.9281
      60      3025      125             0.8841    0.0053     0.8733    0.8941
crpGT3=1
      12      1137       39             0.9668    0.0052     0.9549    0.9756
      24      1089       48             0.9260    0.0076     0.9094    0.9396
```

```
      36      1036       53              0.8809    0.0095     0.8609    0.8981
      48       993       43              0.8443    0.0106     0.8222    0.8638
      60       788       46              0.7998    0.0119     0.7752    0.8219
Note:  survivor function is calculated over full data and evaluated at
       indicated times; it is not calculated from aggregates shown at left.
```

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

```
. sts test crpGT3

         failure _d:  death
   analysis time _t:  obsmos


Log-rank test for equality of survivor functions

       |   Events         Events
crpGT3 |  observed       expected
0      |       760         871.58
1      |       349         237.42
Total  |      1109        1109.00

            chi2(1) =      66.82
            Pr>chi2 =     0.0000

. stcox crpGT3, robust

         failure _d:  death
   analysis time _t:  obsmos

Cox regression -- Breslow method for ties

No. of subjects      =         4933                Number of obs   =      4933
No. of failures      =         1109
Time at risk         =  385305.1317
                                                   Wald chi2(1)    =     64.56
Log pseudolikelihood =    -9145.524                Prob > chi2     =    0.0000

             |               Robust
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
      crpGT3 |   1.687192   .1098327     8.04   0.000     1.485091    1.916796
```

```
. stcox crpGT3

        failure _d:  death
  analysis time _t:  obsmos

Cox regression -- Breslow method for ties

No. of subjects =         4933                     Number of obs   =      4933
No. of failures =         1109
Time at risk    =  385305.1317
                                                   LR chi2(1)      =     60.77
Log likelihood  =    -9145.524                     Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      crpGT3 |   1.687192   .1091937     8.08   0.000     1.486194    1.915374
------------------------------------------------------------------------------
```