EPI 203: Epidemiologic Methods **Name: _____**
Problem Set 2: Disease Occurrence I

Due: 9/29/15 at 1:30 pm section
Possible points: 34

*Note: For the questions marked **"FOR DISCUSSION IN SECTION ONLY:",** please familiarize yourself with these items before section such that we can have a rewarding discussion. Also, please be sure to complete all of the required problem set questions. The first "FOR DISCUSSION" question (question 9c) does not mark the end of the required problem set.*

1. State the measure of disease occurrence (e.g., point prevalence, cumulative incidence, etc.) that is given or can be calculated from the information provided for each study.

(a) Investigators performed a one-time survey of 500 residents of California and asked whether respondents currently had a "common cold". Eighty residents responded "yes". The survey was conducted from January 1, 1997 to June 30, 1998. (1 pt)

(b) Investigators performed a one-time survey of 500 residents of California and asked whether respondents had experienced a "common cold" anytime during the past year. Two hundred residents responded "yes". The survey was conducted from Jan 1, 1999 to Dec. 31, 1999. (1 pt)

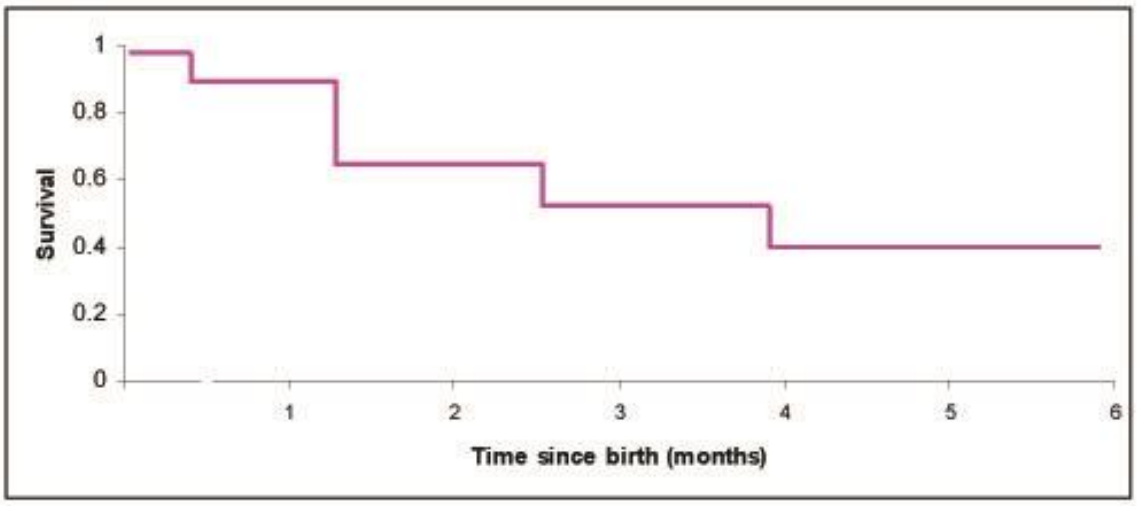
(c) Among 600 initially HIV-negative (uninfected) men, 85 men acquired HIV infection during follow-up. There were no competing events. Because testing was anonymous, it is not known exactly which men acquired HIV. The mean follow-up time (defined as time from entry until either HIV acquisition, drop out, or administrative censoring) for the 600 men was 6.5 years. (1 pt)

(d) The probability that an elderly participant in the Framingham Heart Disease study, which followed residents of Framingham, Mass. for many years for heart disease outcomes, experienced death after 10 years of follow-up was estimated at 0.12. (1 pt)

(e) Among 125 healthy attendees of a church picnic, all were contacted 5 days after the event and 34 reported to have developed diarrhea since the picnic. (1 pt)

***NOTE:** If you are unable to view the graphs in questions 2 and 3 (in the MS Word version of this problem set), please use the PDF version of Problem Set 2 (available on the EPI 203 Syllabus website) to view the graphs.*

2. Results of Kaplan-Meier estimation are shown below from a cohort study of infants with very low birth weight. Infants were followed for death for 6 months following birth.



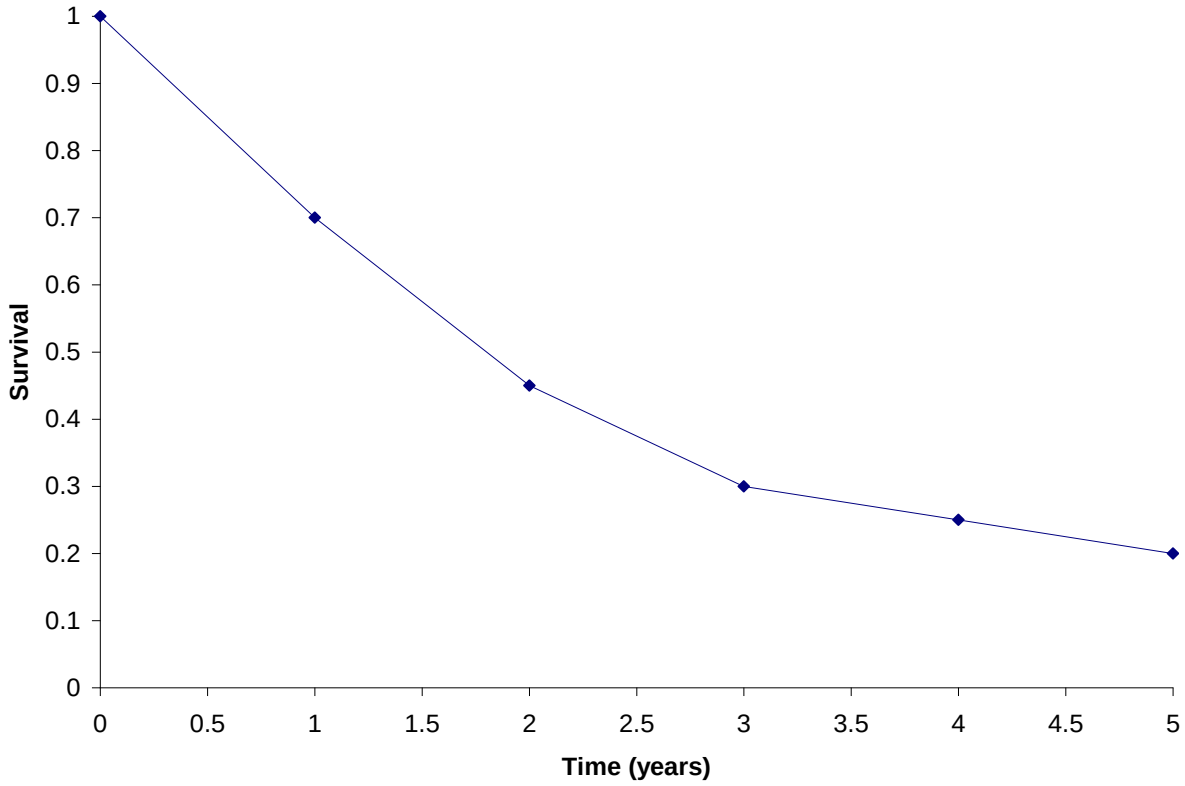
(a) What is the cumulative survival and cumulative incidence of death at 2 months? How about at the end of follow-up? (2 pts)

(b) What is the cumulative incidence of death at 8 months following birth? (1 pt)

(c) At what timepoint does the cumulative incidence of death become 50%? (1 pt)

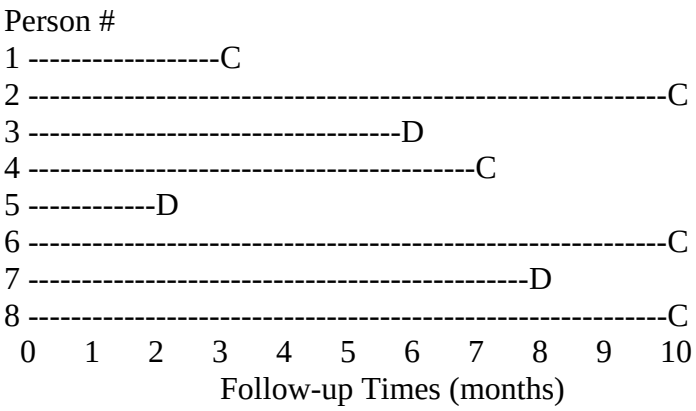
(d) Why does the Kaplan-Meier plot behave like a step function? In other words, why does it look like it does, with a series of steps? (1 pt)

3. Results from a cohort study of cancer patients are shown below. The 5-year cumulative incidence of death was 80%. Does this analysis use the life-table or Kaplan-Meier approach to estimation of incidence? Explain your answer. (1 pt)



4. Using the Kaplan-Meier method of estimating cumulative incidence:

(a) Display in a table, like Table 2-3 in Szklo and Nieto (p.54), the data and the probabilities concerning death and survival for the information below on follow-up of 8 persons diagnosed with lung cancer. In the figure below, C indicates censored and D indicates death. (2 pts)



(b) Calculate the cumulative incidence of death at 8 months. (1 pt)

(c) Draw a graph of the survival function (i.e., survival over time) – by hand (or use the manual drawing tools in MS Word). Please write neatly such that we can accurately evaluate your plot. Be sure that axes are labeled. (1 pt)

5. Consider the following cohort study investigating the occurrence of death following diagnosis of pancreatic cancer in 9 patients.

Patient	Outcome
1	Died at 1 month
2	Still alive at 10 months
3	Died at 6 months
4	Moved away at 7 months
5	Died at 5 months
6	Last known to be alive at 9 months
7	Died at 4 months
8	Stopped coming back to clinic visits after 2 months
9	Died at 10 months

(a) In preparation for a survival analysis estimating the cumulative incidence of death, enter the data into Stata and show a listing of the entire dataset. Show ALL your commands/output. (1 pt)

(b) Using Stata, determine the 10 month cumulative incidence of death. Show your commands/output. (1 pt)

6. The San Francisco Men's Health Study enrolled 394 HIV-infected homosexual men in 1984 and followed them for the next 10 years to determine the natural history of HIV infection (prior to the current era where effective therapy is available). Follow-up occurred by means of examination every 3 months at the central research clinic. The final research study visits were conducted after 9.75 years of follow-up. The outcome was diagnosis of AIDS. Using Stata and the Stata data file that you can download from the course website, called "Epi203survival.dta":

(a) List the Kaplan-Meier estimates of the survivor function (i.e. probability of not experiencing AIDS) at different time points over the 10 year period. Show your commands/output. (Assume no competing events) (2 pts)

(b) Determine the cumulative proportion of HIV-infected persons developing AIDS (i.e., cumulative incidence of AIDS) at 8 years of follow-up. Give your answer to 4 significant digits. (1 pt)

(c) Plot the Kaplan-Meier survival curve for all participants using Stata. Include (1) a table under the plot showing the subjects still at risk; and (2) the 95% confidence interval around each estimate on the curve. Show ALL of your commands/output and make sure to label your axes appropriately. Why is it useful to include such a table of subjects still at risk? (1 pt)

(Hint: Use the drop down menu for the graph command, found at: Statistics/Survival Analysis/Graphs/Kaplan-Meier Survival Function, to locate the options to show subjects still at risk, the confidence intervals, as well as a variety of other options for the Kaplan-Meier graph. Or, search the online user's manual for the options associated with the "`sts graph`" command.)

(d) Plot the Kaplan-Meier failure curve (i.e., incidence of AIDS) and in this plot show:
-- the 99% confidence interval around the points
-- only the first five years of observation, and
-- to emphasize the high incidence of AIDS, limit the y axis to between 0 and 0.5. (1 pt)

7. Stata will show the numbers and pattern of censoring in a Kaplan-Meier analysis if you use the command `sts graph, lost`.

(a) Using the same Stata dataset (Epi203survival.dta), execute this command and display the graph. The number censored at different times of follow-up are displayed on the graph. Assuming that all the subjects started follow-up on the same day (January 5, 1984), is the censoring shown on the plot due to drop out, administrative censoring, or both? (1 pt)

(b) Inspect the pattern of censoring. Assuming again that all the subjects started on the same day (January 5, 1984), are you concerned the nominal estimate of cumulative incidence of AIDS at 10 years might be wrong (i.e., biased)? If so, does the censoring pattern suggest anything about the possible direction of bias in the estimated cumulative incidence of AIDS at 10 years? (1 pt)

8. Let's assume that formal funding for the San Francisco Men's Health Study ended at 10 years such that the subjects were no longer brought in for serial evaluations. The investigators, however, continued to occasionally hear from the research subjects and their families, usually upon the occurrence of death of one of the participants from AIDS. They jotted down these communications and the AIDS diagnoses whenever they happened. At 15 years following start of the study, they decided to calculate another cumulative incidence of AIDS (extending their prior work by 5 more years) based upon the data they had collected by these various communications since the formal study ended. Do you think this updated 15-year estimate is valid? Why or why not? (1 pt)

9. Consider the following abstract from a study of the risk of type 1 diabetes in siblings of type 1 diabetic patients:

   The aims of our analysis were to obtain the risk estimates for type 1 diabetes in the siblings of a Finnish population-based group of childhood-onset diabetic patients and search for demographic and other factors predicting the risk of type 1 diabetes in siblings. We determined the diabetes status of all siblings of all children who are included in the nationwide registry of Finnish cases for whom type 1 diabetes was diagnosed before age 18 years between 1970 and 1979. Siblings' diabetes status from their birth onwards was ascertained by a medical record search through 2001. The total number of person-years during the follow-up of the siblings was 405,685. Of the 10,168 siblings at risk, 647 (6.4%) had been diagnosed with type 1 diabetes by 2001. The cumulative incidence of type 1 diabetes by ages 10, 20, 30, 40, and 50 years in all siblings was 1.5, 4.1, 5.5, 6.4, and 6.9%, respectively.

   Notes: Cumulative incidence was calculated using the Kaplan-Meier technique. Although a few deaths from non-diabetes causes occurred among siblings, consider these negligible in number and for Questions 9a-c, assume no competing events.

(a) The abstract states that 647 (6.4%) of the siblings at risk developed type 1 diabetes. Interpret what this 6.4% means? Is it a useful estimate of incidence? (2 pts)

(b) Based on the information provided in the abstract, would a sibling born in 1991 of one of the childhood-onset diabetic patients contribute data to the estimation of cumulative incidence of diabetes by age 10 years? Would a sibling born in 1991 contribute to the estimation of cumulative incidence of diabetes by age 50 years? What is the earliest possible year of birth for a sibling in this analysis? (2 pts)

(c) **FOR DISCUSSION IN SECTION ONLY:**

The cumulative incidence of type I diabetes by age 20, among siblings of individuals with type 1 diabetes, is stated to be 4.1% in this article. However, in a survey taken in 2001 among 20 year-old siblings of individuals with type 1 diabetes in this region, the prevalence of type 1 diabetes was 12%. What could explain the discrepancy in these reports (4.1% vs. 12%), other than chance? Assume that the method of ascertainment of type 1 diabetes in each report was the same and that it was both highly sensitive and specific. (Clue: we are not looking for a specific biological explanation but rather a general phenomenon).

10. The following abstract and results are from a study of nursing home placements:

> **Objective:** To assess cumulative incidence of and non-cognitive factors predicting nursing home placement in a defined older population.
> **Design and setting:** Six-year follow-up of a population-based cohort living in New South Wales.
> **Participants:** 3654 non-institutionalized residents aged 49 years or older (82.4% of those eligible) participated in baseline examinations during 1992 to 1994.
> **Main outcome measures:** Permanent nursing home admission for long-term institutionalized aged care in New South Wales, confirmed by records of approvals by the regional Aged Care Assessment Team and subsidy payments by government.
> **Results:** After excluding 384 participants who moved from the area or were lost to follow-up, 162 participants (5.0%) had been admitted to nursing homes on a permanent basis by October 1999. Six-year cumulative incidences for nursing home placement were 0.7%, 1.1%, 2.4%, 3.9%, 9.0%, 18.3% and 34.9% for people aged 55-59, 60-64, 65-69, 70-74, 75-79, 80-84 and 85 years or older, respectively. …
> **Conclusions:** Incidence of institutional aged care doubled for each five-year interval from the age of 60 years.

(a) The authors excluded from the analysis the 384 participants who moved from the area or were lost to follow-up. This means they were taken out of the analysis dataset. Do you agree with this exclusion? Explain your answer. (1 pt)

(b) Instead of excluding these 384 participants, let's assume that the authors censored them at their known time of observation and calculated cumulative incidence by the Kaplan-Meier technique. Describe the possible effect of this censoring on the reported cumulative incidence estimates. First, describe a scenario which would cause the reported cumulative incidence to be an underestimate of truth. Second, describe a scenario which would cause the reported cumulative incidence to be an overestimate of truth. (1 pt)

(c) **FOR DISCUSSION IN SECTION ONLY:**
If the cumulative incidence estimates were calculated by the Kaplan-Meier technique, is there anything else – other than the moves/losses to follow-up mentioned in (a) – that makes you concerned about the validity of the estimates? Explain your answer.

11. For the following questions, please refer to the article by Ondrusova and Ondrus to be discussed in Journal Club. (Ondrusova and Ondrus. *Epidemiology and treatment delay in testicular cancer: a retrospective study*. Int Urol Nephrol. 40: 143-148, 2008)

(a) What is the incidence of testicular cancer in 30 to 34 year men in the Slovak Republic between 1993 and 2002? Is the estimate provided an example of cumulative incidence or an incidence rate? (1 pt)

(b) Explain how you think the incidence of testicular cancer in 30 to 34 year old men was derived? In other words, how do you think the authors got their estimate? (2 pts)

(c) **FOR DISCUSSION IN SECTION ONLY:**
The authors contend that the incidence of testicular cancer has increased from 1968 to 2002 in the Slovak Republic. Can you describe any alternative explanations for their data?