Methods in Epidemiologic Research

Sample Problems

Chapter 4 – Measures of Disease Frequency

Preparation

In Chapter 19 data on factors affecting survival after admission to hospital with myocardial infarction are analysed using survival analysis methods. For the following exercises, we will focus on survival over a 5 year period following admission.

The first thing you will have to do is generate three new variables.

- Create a variable indicating the amount of time that the person was under observation. This will be 5 years if the person survived for the full period. If they died, or were lost to follow up (ie "censored") before the end of the 5 year period, it will be the time until death of censoring. When computing this variable, take the following 2 issues into consideration.
  - Before computing this variable, add 1 to the original survival time variable (-surv_mi-) because people who died on the day of admission were listed as having a "survival" of 0 days, but be want them to have a "time under observation" of 1 day.
  - Convert the time from days to years (by dividing by 365.25). This is just to avoid having very small values when estimating rates.

If you are using Stata, some code to generate the variable is

```
gen obs_5yr=(surv_mi+1)/365.25
replace obs_5yr=5 if obs_5yr>5
```

- Create a 0/1 variable which indicates whether or not the individual died during the 5 years following admission.

If you are using Stata, some code to generate the variable is:

```
gen died_5yr=0
replace died_5yr=1 if died==1 & obs_5yr<5
```

- Generate a 0/1 variable indicating whether or not the person survived the full 5 year period.

If you are using Stata, some code to generate the variable is:

```
gen obs_full5yr=obs_5yr==5
```

Questions

1. First determine how many individuals died, survived or were censored (withdrawals) during the 5 year period.

```
.  tab obs_full5yr died_5yr

obs_full5y |       died_5yr
         r |         0          1 |     Total
-----------+----------------------+----------
         0 |       528      1,406 |     1,934
         1 |     1,031          0 |     1,031
```

```
-----------+---------------------+----------
     Total |     1,559      1,406 |     2,965
```

There were 1406 deaths, 528 censored observations and 1031 individuals followed for the full period (not dying)

2. Estimate the 5-year risk of death in 2 ways.

   (a) First, assume (incorrectly) that all individuals were followed for the full 5 year period (ie no withdrawals).

```
. tab died_5yr

   died_5yr |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,559       52.58       52.58
          1 |      1,406       47.42      100.00
------------+-----------------------------------
      Total |      2,965      100.00
```

   It appeared that 47.4% of individuals died during the 5 year period.

   (b) Now adjust the estimate to take into account that there were quite a few withdrawals.

```
. display 1406/(2965-0.5*528)
.52054795
```

   We subtracted ½ the number of withdrawals from the denominator (based on the assumption that withdrawals happened, on average, at 2.5 years). We now estimate that 52.1% of individuals would die during the 5 year period.

   (c) Discuss whether or not this is a good estimate of the 5-year risk of death.

   This is a rather crude estimate of risk because of the quite large number of withdrawals (n=528). Risk can best be estimated directly when the population is closed (no additions and very few withdrawals).

3. Estimate the annual incidence rate of deaths in two ways.

   (a) First, do an approximate calculation based on the assumption that withdrawals took place ½ way through the follow up period.

```
. display 1406/((2965 - 0.5*(1406+528))*5)
.14074074
```

   We have subtracted ½ the number of withdrawals (528) and also ½ the number of cases (1406) on the assumption that they happened, on average, ½ way through the study period. The cases were included in this adjustment because once the event occurred, the individual no longer contributed time at risk (you can only die once). The result would suggest that the population experienced 14.1 deaths per 100 people per year.

   (b) Second, perform an exact calculation based on each person's actual time under observation.

```
. * compute total time at risk for all people (mean * # observations)
.  sum obs_5yr

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
```

```
    obs_5yr |      2965    2.980713    1.945953   .0027379          5
```

```
. display 2.980713*2965
8837.814
. * there were 8837.8 years of observation time
. * compute exact incidence rate
. display 1406/8837.8
.15908937
```

The exact calculation suggests that the mortality rate is 15.9 deaths per 100 people per year.

(c) Discuss whether or not this is a good estimate of the annual incidence rate. Is this estimate a good way to express the frequency of deaths of people in this population? If not, what is its main limitation?

The estimate is fine, but an average annual incidence rate derived from 5 years of follow-up is not a good way to express the frequency with which deaths are occurring in the population, because it assumes that the average value (computed over the full 5 year period) is reasonably constant. In fact, in this population, the mortality rate is very high in the period shortly after admission and then falls off substantially. Consequently, the "average" value may not correctly reflect that mortality rate at any given point in time.

If you repeat the calculations using only the data from the first year after admission, you get an estimated annual incidence rate of 32.9 deaths per 100 people (data not shown). This reflects the much higher mortality rate immediately after the heart attack.

4. Estimate both the 1-year and 5-year risks of death based on the incidence rate computed above.

1-year risk is:

```
. display 1-exp(-0.1591)
.14708894
```

The risk of dying within 1 year of admission for a heart attack was 14.7%

5-year risk is:

```
. display 1-exp(-5*0.1591)
.5486445
```

The risk of dying within 5 years of admission for a heart attack was 54.9%

Note: if you are using Stata, a simpler way to compute risks and rates is using the -cs- and -ir- commands which are designed to compare risks/rates across 2 groups, but which also give you an overall risk/rate.

```
. cs died_5yr sex       /* crude est. of 5-yr risk */
. ir died_5y sex obs_5yr    /* exact est. of annual rate */
```

5. It makes no sense to talk about the "prevalence of death" so instead, compute the prevalence of obesity (defined as bmi>30) among patients admitted to hospital.

First generate an indicator (0/1) variable for whether or not people were classified as obese

```
.         gen obese=bmi>30 if !missing(bmi)
.         tab obese
      obese |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,889       69.83       69.83
```

```
          1 |        816       30.17      100.00
------------+-----------------------------------
      Total |      2,705      100.00
```

30.2% of all admissions were obese.

6. Compute the standard error (SE) and confidence interval (CI) for the crude 5-year risk estimate from 2(a). Compute both approximate and exact estimates. Are they similar? If so, why? If not, why not?

An approximate SE and CI is as follows. (This is a Wald-type CI commonly used for continuous variables – based on the mean and estimated SE)

```
. ci died_5yr
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    died_5yr |       2965     .474199    .0091717        .4562154    .4921826
```

An exact SE and CI is:

```
. ci died_5yr, binomial
                                                      -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    died_5yr |       2965     .474199    .0091702        .4560901    .4923589
```

The two sets of values are very close because: (a) the proportion (.474) is close to .5 (values would differ more if the proportion was close to 0 or 1), and (b) the sample size is large. Note: approximate confidence intervals (based on mean $\pm$ 1.96 SE) may produce end values that are outside the 0 – 1 range.

7. Compute the SE and CI for the annual incidence rate estimate from 3(b).

```
. ci died_5yr, poisson exposure(obs_5yr)
                                                      -- Poisson  Exact --
    Variable |   Exposure        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    died_5yr |   8837.815    .1590891    .0042428        .1508811    .1676275
```

The estimated incidence rate was 15.9 cases per 100 person-years (as seen above), and the 95% CI was from 15.1 to 16.7 cases per 100 person-years.

8. We will now compare the incidence rates across hospitals. However, first create a 4-level categorical variable for -age- with age divided into 4 groups (0-59.99, 60-74.99, 75-84.99 and >=85).

```
. egen age_c4=cut(age), at(0 60 75 85 9999) icodes
```

(a) Determine if the distribution of age categories appears to vary across hospitals (ie do some hospitals tend to see younger or older patients than other hospitals).

```
.  tab hosp age_c4 , row chi
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
  hospital |                  age_c4
```

```
        id |         0          1          2          3 |     Total
-----------+--------------------------------------------+----------
         1 |       321        267        188        114 |       890
           |     36.07      30.00      21.12      12.81 |    100.00
-----------+--------------------------------------------+----------
         2 |        97        157        185        197 |       636
           |     15.25      24.69      29.09      30.97 |    100.00
-----------+--------------------------------------------+----------
         3 |       221        323        368        227 |     1,139
           |     19.40      28.36      32.31      19.93 |    100.00
-----------+--------------------------------------------+----------
         4 |         2          6          4          3 |        15
           |     13.33      40.00      26.67      20.00 |    100.00
-----------+--------------------------------------------+----------
         5 |         3         12         17         18 |        50
           |      6.00      24.00      34.00      36.00 |    100.00
-----------+--------------------------------------------+----------
         6 |         6          6         22         23 |        57
           |     10.53      10.53      38.60      40.35 |    100.00
-----------+--------------------------------------------+----------
         7 |        13          9         21         15 |        58
           |     22.41      15.52      36.21      25.86 |    100.00
-----------+--------------------------------------------+----------
         8 |         9          5          1          1 |        16
           |     56.25      31.25       6.25       6.25 |    100.00
-----------+--------------------------------------------+----------
         9 |         4          6          5          4 |        19
           |     21.05      31.58      26.32      21.05 |    100.00
-----------+--------------------------------------------+----------
        10 |         2          5          3          4 |        14
           |     14.29      35.71      21.43      28.57 |    100.00
-----------+--------------------------------------------+----------
     Total |       678        796        814        606 |     2,894
           |     23.43      27.51      28.13      20.94 |    100.00

          Pearson chi2(27) = 228.1979   Pr = 0.000
```

There appeared to be a substantial difference in the age distribution of people admitted to the 10 hospitals.

(b) Determine if the crude 5-year risk of death (ignoring withdrawals) appears to differ across hospitals.

```
. tab hosp died_5yr, row chi
+----------------+
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
  hospital |       died_5yr
        id |         0          1 |     Total
-----------+----------------------+----------
         1 |       573        339 |       912
           |     62.83      37.17 |    100.00
-----------+----------------------+----------
         2 |       261        392 |       653
           |     39.97      60.03 |    100.00
```

```
-----------+----------------------+----------
         3 |          625         540 |     1,165
           |        53.65       46.35 |    100.00
-----------+----------------------+----------
         4 |            6           9 |        15
           |        40.00       60.00 |    100.00
-----------+----------------------+----------
         5 |           18          36 |        54
           |        33.33       66.67 |    100.00
-----------+----------------------+----------
         6 |           18          40 |        58
           |        31.03       68.97 |    100.00
-----------+----------------------+----------
         7 |           23          35 |        58
           |        39.66       60.34 |    100.00
-----------+----------------------+----------
         8 |           15           1 |        16
           |        93.75        6.25 |    100.00
-----------+----------------------+----------
         9 |           14           6 |        20
           |        70.00       30.00 |    100.00
-----------+----------------------+----------
        10 |            6           8 |        14
           |        42.86       57.14 |    100.00
-----------+----------------------+----------
     Total |        1,559       1,406 |     2,965
           |        52.58       47.42 |    100.00

          Pearson chi2(9) = 118.1041   Pr = 0.000
```

There also appears to be a substantial difference in the 5-year death risks across hospitals.

(c) Compute age-standardized risks for the 10 hospitals. Is there more, or less, variation between the crude or age-standardized risks?

The following block of Stata code does the required calculations to generate age-standardized risks.

```
. preserve
.    keep if !missing(age_c4)
.    collapse (sum) died_5yr (count) n=died_5yr, by(hosp age_c4)
.         format died_5yr n %6.0f
. * compute direct standardized rates
.    dstdize died_5yr n age_c4, by(hosp)
.         * generate a list of the collapsed data
.             list, sep(4)
. restore
```

The summary results of these calculations are:

```
Summary of Study Populations:
     hosp             N      Crude     Adj_Rate       Confidence Interval
 ----------------------------------------------------------------------------
        1           890   0.370787     0.443142    [  0.411640,    0.474645]
        2           636   0.608491     0.546692    [  0.511747,    0.581636]
        3          1139   0.461809     0.449207    [  0.423591,    0.474823]
        4            15   0.600000     0.605218    [  0.349899,    0.860536]
        5            50   0.660000     0.588317    [  0.423716,    0.752917]
        6            57   0.701754     0.596276    [  0.456735,    0.735818]
        7            58   0.603448     0.567175    [  0.454140,    0.680210]
```

```
 8          16   0.062500     0.209399    [  0.209399,    0.209399]
 9          19   0.315789     0.315400    [  0.144488,    0.486311]
10          14   0.571429     0.468187    [  0.277237,    0.659137]
```

As you can see, the difference among the hospitals is greatly reduced from a range of 6% to 70% to a range of 21% to 61% by the standardization process. Clearly, some of the differences among hospitals was due to the difference in age profile of their patients.