**960-584– Biostatistics I– Fall, 2003**

## Homework 1 Solutions, 22 Sept 2002

1. Retrieve the data set from the paper Interpretation of a "Leukemia Trial Stopped Early" by Scott S. Emerson, Phillip L. C.Banks. The data may be found at `http://lib.stat.cmu.edu/datasets/csb/ch14.dat`, on the Statlib server at CMU. The files `http://lib.stat.cmu.edu/datasets/csb/ch14.txt` and `http://lib.stat.cmu.edu/datasets/csb/ch14.sas` have a description of fields in the data set, and a SAS data step to read the data, respectively. If you will use something other than SAS to do computations, note that missing values in the set are marked by a period. Of interest are four variables: Sex, date of entry to study, date of exit from study, and status (alive or dead).

a. Estimate the rate of death in this cohort per person year. Estimate this in such a way that it is allowed to vary by sex. You don't need to make it depend on age (or a proxy), and hence you don't have to go to a lot of trouble to assign events and time at risk to different age groups.

*Here is the sas code:*

```
data emerson; set emerson;
  case=0; if b3="D" then case=1;
  time=lastt-start; run;
proc sort data=emerson; by sex; run;
proc means data=emerson sum; output out=sums sum=; by sex; run;
data sums; set sums; rate=case/time; run;
proc print data=sums; var sex case time rate; run;
```

*and here are the resulting rates (in case/day):*

| OBS | SEX | CASE | TIME | RATE(per day) | RATE (per yr) |
|---|---|---|---|---|---|
| 1 | F | 39 | 34801 | .0011207 | .409 |
| 2 | M | 48 | 29868 | .0016071 | .587 |

b. Is standardization necessary for comparisons across sex? Why or why not?

*No, because the end point is all mortality, and the difficulty arising from deaths from other causes does not arise.*

2. Waller, Turnbull, Clark, and Nasca present data containing numbers of new cases of lukemia in 791 census tracts between 1978 and 1982 (five years in all), the 1980 populations of these tracts, and the centers of the tracts. They also provide locations of eleven superfund toxic waste sites. The data can be found at `http://lib.stat.cmu.edu/datasets/csb/ch1a.dat` and `http://lib.stat.cmu.edu/datasets/csb/ch1b.dat` . Consider the derived data set consisting of the populations, lukemia cases, and distances from the nearest superfund site, found at `http://stat.rutgers.edu/~kolassa/960-584/lukemia.dat`. (The numbers of cases here are not integers, since cases near the boundaries of tracts were divided among nearby census tracts). These distances are rounded into discrete groups.

a. Calculate the lukemia rates for each of the distance groups. Use as an estimate of person years at risk the 1980 population times the number of years under observation.

*We'll calculate person years at risk, and then collapse across distance groups:*

```
data lukemia; infile 'lukemia.dat'; input pop case dist;
  pyar=5*pop; run;
proc sort data=lukemia; by dist; run;
proc means sum data=lukemia noprint; by dist;
  output out=sumset sum=; run;
data sumset; set sumset; rate=case/pyar;
  cil=rate*exp(-1.96/case); ciu=rate*exp(1.96/case); run;
proc print data=sumset noobs;
  var dist case pyar rate cil ciu; run;
```

*and here are the results:*

```
dist  case    pyar      rate      cil       ciu
  0  228.20 1505730  .000151554 .000150258   .000152862
 15   87.82  887305  .000098974 .000096789   .000101208
 30  162.00 1690725  .000095817 .000094665   .000096983
 45  110.19 1152035  .000095648 .000093962   .000097365
 60    3.79   52570 .000072094 .000042984   .000120920
```

*We see that the highest rates are close to the superfund site, and lower rates are farther away, with the lowest rates at the greatest distance.*

b. Calculate a confidence interval for each of these rates, and for the difference between the group closest to superfund sites and those farthest away. Comment on your findings. *Each of these rates is equivalent to an SMR if the standard expected value is 1. Hence we can use the same formula for the confidence interval as we did for the SMR: CI= $\hat{\lambda}\exp(\pm 1.96/\sqrt{d})$ , where $d$ is the number of cases. The calculations are shown as part of the results for part a. Ignore for the present problems of multiple comparisons. The closest group clearly has a higher rate than the others. The remainder are statistically indistinguisable. The group farthest away contains too few individuals to accurately distinguish it from its nearer neighbors.*

*I also asked for the difference between the rates for the closest and the farthest away groups. Recall that the variance for each rate is estimated as the number of cases divided by the person years at risk. Hence the variance of the difference is the sum of the variances, and so the confidence interval for the rate difference is* $.000121244 - .000057675 \pm 1.96 * \sqrt{228.20/1505730^2 + 3.79/52570^2}$ . *You could do this calculation in SAS using*

```
data sumset; set sumset; mult=0; if dist=0 then mult=-1;
  if dist=60 then mult=1; rd=rate*mult;
  var=mult**2*case/pyar**2; run;
proc means data=sumset sum noprint ; var rd var;
  output out=last sum=; run;
data last; set last;
  cil=rd-1.96*sqrt(var); ciu=rd+1.96*sqrt(var) ; run;
proc print data=last noobs; var rd var cil ciu; run;
```

*and get*

```
 rd       var       cil          ciu
-.00007946   1.472E-9   -.00015466   -.000004260
```

*Hence the confidence interval is* $(-.00015466, -.000004260)$ .