Homework 6 Solutions, 8 Dec 2003

1. 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. Calculate a confidence interval for the relative risk of the second-closest group relative to the closes group. Let λ_1 be the lukemia rate in the closest group of census tracts, let m be the number of census tracts closest to the superfund sites, let W_i be the square roots of the number of lukemia cases, and let Q_i be the person-years at risk. Let λ_2 be the lukemia rate in the next closest group of census tracts, let n be the number of census tracts in the next closest group, let Y_i be the square roots of the number of lukemia cases, and let Q_i^* be the person-years at risk. Then $\mathrm{E}[W_i] = \sqrt{\lambda_1 Q_i}$ and $\mathrm{E}[Y_i] = \sqrt{\lambda_2 Q_i^*}$. Hence $\lambda_2/\lambda_1 = [(\sum_i \sqrt{Q_i}/m)/(\sum_i \sqrt{Q_i^*}/n)]^2 [\mathrm{E}\left[\bar{Y}\right]/\mathrm{E}\left[\bar{W}\right]]^2$.

   a. Use Fieller's method to get a confidence interval for $\mathrm{E}\left[\bar{Y}\right]/\mathrm{E}\left[\bar{W}\right]$, and transform it to get a confidence interval for λ_2/λ_1 .

The SAS code

```
data lukemia; infile 'lukemia.dat';
  input pop case dist; pyar=5*pop;
  rtc=sqrt(case); rtpyar=sqrt(pyar);
  if dist>20 then delete; run;
proc sort data=lukemia; by dist; run;
proc means data=lukemia n mean stddev;
  var rtc rtpyar; by dist; run;
```

   produces the summaries

```
------------------ dist=0 --------------------
Variable    N        Mean      Std Dev
------------------------------------------------
rtc       130     1.1335647    0.6885219
rtpyar    130   100.4756057   38.7133018
------------------------------------------------

------------------ dist=15 ------------------
Variable    N        Mean      Std Dev
------------------------------------------------
rtc       124     0.6453722    0.5423027
rtpyar    124    79.3259361   29.4974280
------------------------------------------------
```

To obtain a confidence interval for $\mathrm{E}\left[Y_j\right]/\mathrm{E}\left[W_j\right]$ *, then* $\bar{Y} = 0.6453$ *,* $\sigma = 0.5$ *,* $\bar{W} = 1.1336$ *,* $\tau = 0.5$ *,* $n = 124$ *,* $m = 130$ *, and* $\rho = 0$ *. Then the CI is* $(0.4846, 0.6608)$ *. One uses the normal distribution critical value. To get the final result, square the endpoints, and then multiply by the ratio of the average means squared. This result is* $(0.37677, 0.70052)$ *. The following SAS commands might have been used to do these calculations, and the calculations for the next part.*

```
data temp; set temp; lmrtpyar=-log(rtpyar)*(-1)**_n_; run;
proc means data=temp sum noprint; var lmrtpyar; output out=temp1 sum=;
   var lmrtpyar; run;
data temp1; set temp1; rrtpyar=exp(2*lmrtpyar); output; output; run;
%include 'biostat1macros.sas';
data coef; input
barY   barW m n sigma  tau   df rho; cards;
0.6453 1.1336 130 124 .5 .5 . 0
;
run;
data out; run;
%doci(coef,.05,'Part a');
data coef; set coef; sigma=0.5423; tau=0.6885; run;
%doci(coef,.05,'Part b');
data out; set out; if name='Known Var.' then; else delete; run;
data out; merge out temp1; keep runn upper lower nlower nupper rrtpyar;
   nlower=lower**2*rrtpyar; nupper=upper**2*rrtpyar; run;
proc print data=out; run;
```

Here the file `biostat1macros.sas` *has the macro doci that I used in class.*

b. Now repeat the previous calculation, without the assumption that the variances of the counts of square roots of lukemia cases are approximately one quarter, and without assuming that they are equal to each other.

In this case, $\sigma = 0.5423$ *, and* $\tau = 0.6885$ *. The resulting CI is* $(0.35852, 0.73951)$ *. Note that even the minimal number of degress of freedom is 123, and hence one may use the normal distribution critical value.*

2.   Ryan, Joiner, and Ryan (1976) present data on volumes of lumber available from 31 black cherry trees. Diameters (D_i), in inches, heights, in feet, and volumes (V_i), in board-feet, are given in a file found at `http://www.statsci.org/data/general/cherry.txt`. For the present ignore heights. The best fit for the model $V_i = \beta_0 + \beta_1 D_i + \epsilon_i$, for ϵ_i independent normal variables with mean zero and variance σ^2 , is given by $\hat{\beta}_0 = -36.9435$ and $\hat{\beta}_1 = 5.0659$. Standard errors are 3.2548 and 0.2393 ; furthermore, the correlation between these -0.9739 . Give a confidence interval for the diameter of trees whose expected volume is 16 board-feet. Note that for models of this form, the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is known without error, and the standard errors for $\hat{\beta}_0$ and $\hat{\beta}_1$ are a fixed multiple of a quantity whose square has a χ^2_{29} distribution.

You weren't asked to calculate the quantities from the regression yourselves, but if you had been, you could have done this:

```
/* Data on 31 black cherry trees, diameters, heights, and volumes,
from Ryan, T.A., Joiner, B.L. and Ryan, B.F. (1976) Minitab
Student Handbook. N. Scituate, MA: Duxbury*/
options ls=78;
data trees; infile 'trees.dat'; input d h v; run;
proc genmod data=trees; model v=d/corrb pscale; run;
```

We need to bound the Δ solving $16 = \beta_0 + \beta_1\Delta$; $\hat{\Delta} = (16 - \hat{\beta}_0)/\hat{\beta}_1$. Use the CI formula with $\bar{Y} = 16 - \hat{\beta}_0$, $\bar{W} = \hat{\beta}_1$, and with the appropriate standard errors for σ and τ. This might be evaluated using

```
data coef; input
barY    barW m n sigma  tau   df rho;
barY=16+barY;
cards;
-36.9435 5.0659 1 1 3.2548 0.2393 29 0.9739
;
run;
data out; run;
%doci(coef,.05,'hi');
proc print data=out; run;
```

to obtain the confidence interval $(9.83519, 10.6454)$. The sign of ρ above was changed, since the sign of β_0 in the numerator of the fraction is negative.