Fall 2003 Course EPIB-626: Risks and Hazards

Questions for Assignments [updated Sept. 15, 2003]

material in www.epi.mcgill.ca/hanley/c626/ unless otherwise specified

(username: c626 ; password: 8 letters, H***J*## both case-sensitive)

1 Comment on the validity of the conclusions regarding scouting injuries [Under **Notes on Poisson Distribution**] . Hint: think person-time!

2 [pp 20-30 helpful for q's 3-6] Read the draft of the Notes on Poisson Distribution and indicate -- on writing on the paper -- items that are not clear and that could be better explained. Hand back the marked up notes (a new set of notes, based on your suggestions / comments / corrections, will be posted on the 626 www page).

3 Calculate a 95% CI for the SIR and perform a test (at alpha = 0.05 2-sided) of SIR=1 for the Alberta Sour Gas Study [restrict attention to the 33 vs. 36.3 [Index Area1970 Cohort Females vs. (1) Southern Alberta excl. Calgary, Lethbridge, & Medicine Hat (RP1)].

Carry out the same inference procedures, but imagine the concerned area or cohort was much smaller .. so that 3 cases were observed where 0.45 were expected.

4 [Optional, but relevant for question on Ralox] Calculate 95% CI's and perform tests (at a = 0.05 2-sided) for the rate ratio and rate difference based on the Rothman & Boice data given on page 28 of Poisson notes.

5 Show how you think the authors got the various numbers in the following passage from the study on DAYLIGHT SAVINGS TIME AND TRAFFIC ACCIDENTS [Under **Notes on Poisson Distribution**]. The loss of one hour's sleep associated with the spring shift to daylight savings time increased the risk of accidents. The Monday immediately after the shift showed a relative risk of 1.086 (95 percent confidence interval, 1.029 to 1.145, x2 = 9.01, 1 df, P<0.01).

Comment on whether the Poisson distribution is actually appropriate in this accidents example. Restrict attention to the Spring. Hint: imagine you had the individual yearly Monday data not just for 1991 and 1992 but for say 1980 to 1999. Do you think the year to year variation in the 20 "pre-time-change" Mondays would be Poisson?

Daylight savings time and traffic accidents [letter] [see comments].
Comment in: N Engl J Med 1996 Aug 1;335(5):355-6; discussion 356-7, Comment in: N Engl J Med 1998 Oct 15;339(16):1167-8 ; Source: New England Journal of Medicine. 334(14):924, 1996 Apr 4.

6 Accidents involving female vs. male airline pilots [Under **Notes on Poisson Distribution**]

7 Questions 1-6 based on Abstract on Raloxifene & Risk of Breast Cancer in Postmenopausal Women [part of final exam from a previous year]

8 Question 2 based on article A CONTROLLED TRIAL OF A HUMAN PAPILLOMAVIRUS TYPE 16 VACCINE (exam question, December 2002) [*similar to worked e.g. on page 29 of Poisson Notes*]

9 Examine the trend in the sex-ratio, and in the proportion of male births, over the last 60 years in Canada [Under **Regression for Risks(proportions) and Rates**] .

10 Fit a relationship between maternal age and parity and the incidence of Down's syndrome [Under

Datasets] . Verify that you get very similar answers by Binomial and Poisson regression. Why?

11 Examine the "hurricane" data (the 10 datapoints from USA Today, under **Datasets**]) for evidence that the rate of hurricanes has been increasing over the past century.

Note that the data for the 1990's decade only cover the 5 years 1990-1994. Try not to (i) double the count to make the amount of experience commensurate with the full decades (if you did this, the estimation procedure would not know that you had "manufactured" data) (ii) drop this last datapoint. Instead, use the familiar "expected number of cases = Rate x Amount of experience" relation to see how to "retain" this last datapoint properly.

12 What is Simpson's paradox? What are your best two (real!!) examples of Simpson's paradox (a) in epidemiology (b) other related fields. In each example, explain how the paradox occurs.

13 Analyze the relationship between cigarette smoking and lung cancer rates in the British Doctors' study (data on 626 web page Under **Datasets**]; if you wish , consult B&D Vol II). Pay attention to modification by age.

14 Read the "Women are Safer Pilots" story and carry out the requested analyses for the rate ratio, both by regression and non-regression methods. Assume that on average, the women pilots fly (i) just as much as the men pilots (ii) half as much. Hint: it is not the absolute, but rather the relative sizes of the "denominators" that you need - make up some reasonably realistic denominators for each sex. [the regression analysis of the Rothman-Boice data (p 14/15 of my regression notes) is a good "template" for the regression analyses]

15 What do you think the mystery data [under **Datasets**] come from? Would a regression analysis help you to figure out the source?

16 Examine the yearly hurricane data (in the file "storms data" with 110 counts, all from the official source) for evidence of trends.

17 Refer to the data in the article on crash rates on a toll highway [and accompanying SAS file, under Regression for Risks(proportions) and Rates].

"The rates (total crashes, last 2 columns) showed significant statistical evidence of heterogeneity" (2nd sentence of findings)

- Using just a hand calculator or spreadsheet, replicate the calculations that led to this p-value (see statistical note in appendix 1 of the article .. i.e. compare the '1 x 7' table of *observed* numbers of total crashes, with the corresponding 1 x 7 table of *expected* numbers under heterogeneity, and calculate Sum(O - E)^2 /E, taking the sum over the 7 years. If you think of each 'mile driven' as the unit of observation, you could set up the observed data as a more traditional ***2*** x 7 table of observed numbers -- with one row being the numbers of 'miles in which there was a crash' and the other the 'number of miles in which there wasn't' and you could do the same for the expected numbers, but you will find that the contribution to the sum of the 14 (O- E)^2/E 's will virtually all come from the 7 in the top row, where the crashes are. [if interested as to why, see my notes on '2 x 2', '2 x 1', 1 x 2' and '1 x 1' tables in page 9 of my notes for chapter 9 of course 607 in 2001].

- [data for all 7 years] Run the null (intercept-only) models using PROC LOGISTIC and using PROC GENMOD (the latter separately with binomial and Poisson errors) .. -- SAS code supplied.

    - Interpret the intercept coefficient in each model, and relate it to the overall rate for the 7 years (i.e., to the 9650 crashes / 7611165 thousand-vehicle-miles.

    - Match up the Goodness-of-Fit statistics with the one you calculated by hand. Comment on the

possible sources of this heterogeneity.

[NB you may be unfamiliar with the 'numerator/denominator' way of specifying the 'y' on the left hand side of the logistic or binomial model.. you probably used for your 681 logistic project a separate line of data for each 'person', so that y = 0 or 1; (by default) the denominator was always 1, so the numerator/denominator form would always be 0/1 or 1/1 . Proc logistic assumes, unless told otherwise, that the denominator is 1 [but it will complain if the numerator is greater than the denominator!] Here the denominator is aggregated .. otherwise, it would be a very large file of 7611165000 separate lines, 1 per mile driven!. the same aggregate format is used for the data on Down's syndrome study [see datasets on web page] ..each line of data corresponds to a *number of women with the same covariate pattern* (defined by age and parity).

- Run the null (intercept-only) Poisson model using PROC GENMOD for the *6 years before the change* -- again given in the SAS code.

  - Interpret the intercept coefficient in this model, and relate it to the overall rate for the 6 years. Again, comment on the Goodness-of-Fit statistics.

- Run the *non-null (linear-change in log-rates)* Poisson model using PROC GENMOD for the *6 years before the change* -- again see supplied SAS code. [note the use of "1968 = year 0" variable]

  - Interpret the coefficients in this model, and relate them to the trend in rates for the 6 years [i.e. plot the observed rates, and draw in the fitted curve. How close to a straight line is the fitted curve? *[the log link specifies that the ln of the rate is linear over calendar time]*

  - Comment on the Goodness-of-Fit statistics.

- Run the *null* and *non-null (linear-change in rates)* <u>*additive*</u> Poisson model using PROC GENMOD for the *6 years before the change* -- again see supplied SAS code. [note the use of "1968 = year 0" variable]

  - Try to interpret the coefficients in this model, and relate them to the trend in rates for the 6 years i.e. plot the observed rates, and draw in the obtained the fitted line. [to do so, it is easier if you first interpret the *null* model *literally*:.. i.e. because of the [forced] zero intercept [specified by the NOINT option] , the model is a simple statement of the most basic of all epidemiologic relationships, namely

    (1) expected # of cases = rate × denominator

    so by writing the simplest of all non-null regression models

    (2) # cases = β × denominator + Poisson variation around expected value,

    the fitted β is the overall observed rate.

    Now it is easier to move on to the non-null model, where the rates go up [or down] linearly over calendar time. ie **rate =** $\beta_0 + \beta_1 \times$ **year.** Substitute this 'rate pattern' into equations 1 and 2, and multiply across by the denominator, yielding the model

    # cases = $\beta_0 \times$ <u>denominator</u> + $\beta_1 \times$ <u>year × denominator</u> + Poisson variation ,

    "Y" = $\beta_0 \times X_0 + \beta_1 \times X_1$ + Poisson variation .

- Comment on the Goodness-of-Fit statistics from this additive model. Why *in this example* are they reasonably close to the ones from the multiplicative [i.e. log rate is linear] model?

18 Question 2 based on article "Universal hepatitis B vaccination and the decreased mortality from fulminant hepatitis in infants in Taiwan", from December 2002.

19 Question based on article ""Impact of a helmet law on two wheel motor vehicle crash mortality in Barcelona" (Injury Prevention 2000; 6: 184-188)", from previous exam .

20 Refer to the data from John Snow's study, given on bottom of column 3 of page 1 of handout for Sept 05 lecture for Med2 [on med2 website, reachable from link at top of 626 website: username med2, password: same as for the cxxx epidemiology courses]. Calculate a 95% CI to accompany the rate ratio of 13.3. Do the same for the ratio estimates based on the denominator series of 100 and 1000 (first column, page 2... [n practice, you would not observe the quasi-denominators shown there, but rather these expected numbers ± some sampling variation].

21 Refer to the article A POPULATION-BASED STUDY OF MEASLES,MUMPS,AND RUBELLA VACCINATION AND AUTISM [under **Nov 11 lecture in med2,** reachable from link at top of 626 website: username med2, password: same as for the cxxx epidemiology courses]. See also page 2 of the handout of Nov 11, dealing with CI's for ratios, and the fact that for ***log-based CI's*** (instead of the usual ± a margin of error for 'regular' statistics) for ***ratios***, one can calculate a "multiplied-by/divided-by" factor in order to arrive at the upper/lower limits. To convince yourself, calculate the CI's for the ratio of 13.1 on page 2, and the 1.44 ratio on page 3, by your usual way, and compare them with the answers from the "multiplied-by/divided-by" method shown.

22 Refer again to the article A POPULATION-BASED STUDY OF MEASLES,MUMPS,AND RUBELLA VACCINATION AND AUTISM [under **Nov 11 lecture in med2,** reachable from link at top of 626 website: username med2, password: same as for the cxxx epidemiology courses] and to the righthand column of page 3 of the med2 handout of Nov 11, dealing with the reason for the big difference between the crude rate ratio of 1.44 and the 'adjusted' rate ratio of 0.92.

• Explain to a journalist why the big difference in this example.

[The link How could one get a crude Rate Ratio (RR) of 1.45? gives a bit more detail]

• Complete the exercise at the bottom of the second page in the link Why can the crude Rate Ratio (RR) be 1.45 if RR=1 at all ages and in all years? Hint: use each vertical slice as a 'stratum' and use the Mantel-Haenszel summary rate ratio (Sum over 8 vertical slices*)

$$\text{rate ratio} = \frac{\#\text{Exposed Cases} / \text{Exposed P-T}}{\#\text{Unexposed Cases} / \text{Unexposed P-T}} = \frac{\#\text{Exposed Cases} \times \text{Unexposed P-T}}{\#\text{Unexposed Cases} \times \text{Exposed P-T}}$$

$$\text{MH summary rate ratio} = \frac{\text{Sum}\,[\#\text{Exposed Cases} \times \text{Unexposed P-T} / \text{Combined P-T}]}{\text{Sum}[\#\text{Unexposed Cases} \times \text{Exposed P-T} / \text{Combined P-T}]}$$

and calculate a (test-based) CI to accompany it. (*ideally, should summarize over all 36 age-year cells)