Department of Epidemiology and Biostatistics, McGill University
Course 513-607 Inferential Statistics in Medicine, May/June 1995
Final Examination, given out June 20, due back June 29.

Refer to the article "**Randomized Controlled Trial of Routine Cervical Examinations in Pregnancy**" by Buekens P. et al in The Lancet Vol 344 pages 841-814 September 24, 1994.

1 (Statistical Methods, paragraph 1)

(i) Verify the sample size requirement of 7000 per group, and the calculation for the reduced sample trial of 3000 per group.

(ii) Given the huge budget and 'not easy to replicate' nature of this trial, one could argue for demanding higher power. Redo the calculations using a beta of 10%.

(iii) For an alpha of 0.05 and a given delta, what is the percentage increase in sample size as one goes from beta=0.2 to beta=0.1?

2 (Statistical Methods, paragraph 2)

(i) Illustrate, using as an example the first outcome in Table 3, that statistical comparisons of the frequencies of binary outcomes yield the same p-value whether carried out with the chi-square test for a 2x2 table [your choice of computation method] or the z-test for 2 proportions. Use enough decimal places to be sure they really are the same.

(ii) Repeat both calculations, but this time using the continuity correction

$$x^2 = \sum \frac{\{|o-e|-0.5\}^2}{e} \quad \text{or} \quad \frac{N\ \{\ |ad - bc\ | - N/2\ \}^2}{r_1\ r_2\ c_1\ c_2} \quad \text{or} \quad \frac{\{\ |\ a - E[a]\ |\ - 0.5\}^2}{r_1\ r_2\ c_1\ c_2\ /\ N^3}$$

$$z = \ = \frac{p_1^* - p_2^*}{\sqrt{p\{1-p\}\{1/n_1 + 1/n_2\}}}$$

where, if p_1 is the larger and p_2 the smaller of the two proportions, $p_1^* = \frac{y_1 - 0.5}{n_1}$ and $p_2^* = \frac{y_2 + 0.5}{n_2}$
(Colton explains this correction in the z-test for proportions on page 165 but then leaves a typographical error in the z formula)

Note a propos the CI's for RRs: The Taylor series CI for the ratio of proportions is described in §15.2.1 of Kleinbaum, Kupper and Morgenstern's text "Epidemiologic Research".

3 (Results: Characteristics of study population)

"The trial groups were similar in age, education level and baseline obstetric history (Table 1)"

(i) One would expect with such large sample sizes that the balance would be excellent; but just how close should the means be? For example, should the n's of ≈2750 and ≈2750 "guarantee*" that the average age in the two groups would not differ by more than 0.1 year or 0.5 years or 1.5 years? Assume some justifiable standard deviation of individual ages and calculate the possibilities for the difference between the average age of one random half of the subjects and that of the other half. Sketch a frequency distribution to illustrate the results of your calculations. Hint: the question concerns the sampling variation of $\bar{y}_1 - \bar{y}_2$ calculated assuming randomization. [* nothing can be "guaranteed" but use as an operational definition "95% sure"]

(ii) Do the same for the difference in the frequency of primapara.

Note: While these types of calculations are the basis for significance tests for outcomes, they should not be used to carry out formal tests of hypotheses on baseline data from RCT's (Table 1 in most clinical trials). There is no great reason to calculate p-values for baseline differences unless one wishes to check if they carried out the ranoomization correctly. A much more important question than whether any imbalances are statistically significant is the magnitude of the differences and how much distortion these imbalances actually make to the comparison i.e. . it's a question of "embarrassing" rather than "statistically significant" differences.

4 (Results: second paragraph)

The authors report a total of 20+0+23+1 = 44 multiple births among 2719+2721 = 5440 women followed up until delivery.

(i) Calculate a 95%CI for the frequency of multiple births per 1000 women. What confidence do you have in this interval estimate as an estimate of the frequency of multiple births in general in these countries?

(ii) The authors calculated preterm rates per 100 total births. They used the chi-square [= z^2] tests to compare them. Are all of the statistical requirements for such tests met? If not, will the reported p-value be too big or too small?

5 (Results: Cervical examinations and interventions)

(i) What is the purpose of reporting and comparing the number of cervical examinations in the two groups? Is it advisable to perform a formal test of significance for this comparison?

(ii) Why are medians rather than means used in the third paragraph?

(iii) From the point of view of budget people, why is the mean more relevant than the median?

(iv) Are t-tests on mean numbers of visits or mean lengths of bed rest stay or hospital stay contra-indicated by your answer to (ii)?

(v) The authors used the Mann-Whitney test (= Wilcoxon Rank Sum Test) for comparisons of "distributions" (third paragraph). If they had consulted you about between parametric vs. non-parametric tests for these, what would you have advised and why?

(vi) How does one obtain a p-value for the Rank Sum test when the n's are so high?

6 (Outcomes)

Throughout, the authors are more concerned with ratios of proportions (what they call risk ratios or RR's) than with differences in proportions (what we might call risk differences or RD's). If we wanted to test that RR=0, using alpha = 0.05 two sided; for the outcomes in Table 3, how could we infer the results of such tests without actually carrying them out?

7 (i) Make a rough plot the 7 CI's for the RR in the top half of Table 4 in the graphical style used in meta-analyses.
Why are the CI's not symmetrical about the point estimate?

(ii) Sketch what a graphical display of the results would look like if presented on a RD rather than an RR scale.

8 (i) Do you agree with the need for the Bonferroni correction (i.e. using alpha = 0.05/14 rather than 0.05) when interpreting the RR's for the different countries? see M& M p 742- for methods for 'correcting for' multiple comparisons; they concentrate on comparisons of several treatments; but the issue can also be raised concerning comparisons of the same two treatments in several subgroups of the same dataset;. The Bonferroni correction (see page 844) involves dividing the overall alpha (in this case 0.05) by the number of tests carried out (in this case 14), and using this stricter alpha for each separate test.

(ii) If one performs 14 independent tests of a null hypothesis using an alpha of 0.05 for each one, and the null hypothesis is indeed true, what is the probability of at least one false rejection among the 14? (as is often the case with calculations involving "at least one", it is easier to calculate it as 1 minus the probability of all 14 test being negative)

(iii) If -- again when Ho is true -- one uses an alpha of 0.05/14 for each test, what is the probability of at least one false positive test? (It will not come out exactly to 0.05 but it will be close)

(iv) Can you make a case why -- even if you agree with the principle of correcting for multiple independent tests -- dividing by 14 in the Bonferroni correction may be overly stringent in this example and why a smaller divisor -- somewhere between 7 and 14 -- might be more reasonable? (Think of the two tests done for each country)

(v) Do you believe that the results in Spain and Portugal could be chance variations and the significant results are a consequence of overtesting (overfishing?) or do you believe they are real?

(vi) How much of your interpretation comes from the fact that these occurred in Spain and Portugal? Would your interpretation have been different if the countries in Table 4 were blinded and referred to only as country 'A' to country 'F'?

(vii) What does you answer say about relying solely on the p-value in judging whether a difference is 'real' or not?

9 (i) What test is appropriate for testing the hypothesis that the frequency of preterm deliveries differs among the countries? Do not carry out the test, but point to a similar example in Moore and McCabe.

(ii) What are the alternative and alternative hypotheses tested?

(ii) If the results were significant, what conclusions could one safely draw?

10 Are the results 'definitely negative' or simply 'inconclusive'? Advocates of repeated cervical examinations will surely point to the fact that the study did not have sufficient power to detect a risk reduction of 20% in the preterm rate i.e. an RR of 0.8.

(i) Calculate the power of a study with 2750 and 2750, with alpha =0.05 two sided as before, to detect such a reduction? Hint: solve for Z_{beta} in the formula linking n, alpha beta and delta.

(ii) Such power calculations are relevant only for planning purposes and are of little relevance after the data are in. Instead, one should use the calculated CI: in this case the 95% CI for the RR was 0.85 to 1.29.

(iii) Suppose you are involved in a panel discussion about the study. One panelist focuses on the inadequate power but does a poor job of communicating in non-technical language what statistical power is. Provide a non-technical but correct 'translation'.

(iv) Another panelist uses the CI from the study rather than the pre-study calculations of power, but again does a poor job of conveying the meaning of a CI. Give an understandable and at the same time technically correct translation of the CI.

(v) A third panelist says "we should all be much more Bayesian about all of this". Again, translate "Bayesian".

11 The trialists considered a risk reduction of 20%, or from 5% to 4%, as clinically significant.

(i) What factors go into deciding what is clinically significant?

To many, what is relevant is the RD rather than the RR, since one can directly calculate the number required to treat in order to prevent one bad outcome as 1/RD. Proponents such as Sackett call this the "Number Required to Treat" (NRT). So if the RD was 1.1% (7.7% minus 6.6%), one would need to intervene on NRT=100/1.1 = 91 pregnancies to prevent one low birthweight birth.

(ii) However the estimate of RD, and thus of NRT, has some sampling variability. Use the 95% CI for RD for low birthweight to find a 95% CI for NRT.

Refer to "**Differences in proximal femur bone density over two centuries**" in The Lancet Vol 341 pages 673-75, 1993.

1 *"The precision of measurement was 1.2% (femoral neck) and 1.70% (Ward's triangle)"* [3rd paragraph of Subjects and Methods]

The authors do not define the term 'precision'. How do you think they defined and calculated it?

2 Put the "slope=0.197" in Table III into plain words.

3 "in the ancient femora, there was no significant loss of bone density premenopausally in either region" [1st half 2nd sentence Results]

(i) Draw in the regression lines for the ancient femora, both regions, premenopausal.

(ii) Do you have enough information to directly (i.e. without having to go back to the raw data points) assess that there was no (statistically) significant loss? If yes, do so; if not, say why not.

4 *"in striking contrast to modern women"* refers to the slope of 0.197 vs. that of –0.658 (neck) and –0.162 cs. –0.921 (triangle).

If you had the regression printouts for the ancient and modern-day groups analyzed separately, how would you use them to verify that this "striking difference" was indeed statistically significant?

5 The answer from the test involving r=0.424 was "*p<0.005"; what question did the test answer?

6 Are you convinced about the numerical differences found? If you are, what do you think the main reasons for these differences are?

[For information only: the SEE is in fact the SD of the residuals. SEE is not a very descriptive term]

Again, as with the MidTerm examination, list the contributions of all collaborators and consultants.