Epidemiology & Biostatistics/McGill University **In Class Exam Course 513-607A** **December 2001**

Instructions

This examination has 13 pages. Please check it as your first step.

Answers are to be written on the question paper, in the spaces provided.

Unless specifically asked to in a particular question, you do not need to complete any detailed calculations. Instead, answer as though you were setting up the computing task for your research assistant to complete.

If there is a possibility of ambiguity, be clear about 1- and 2-sided hypotheses, levels of confidence, degrees of freedom, whether 2 sample or 1 sample procedures, paired or unpaired, which table or formula from your textbook is relevant, etc.

Even though the points below add to 147, the exam will be marked out of 120, so you may skip some items.

Item	Gen.	Hips	Vit. C	Bile Duct	AMI
				vs. age	Twins
1	__/2	__/5	__/3	__/4	__/3
2	__/4	__/2	__/2	__/4	__/3
3	__/3	__/4	__/3	__/3	__/5
4	__/3	__/2	__/2	__/5	__/2
5		__/2	__/2	__/5	__/3
6			__/2	__/5	__/4
7			__/4	__/5	
8			__/2	__/4	
9			__/2	__/10	
10			__/3	__/3	
11			__/4		
12			__/5		
13			__/4		
14			__/4		
15			__/5		
16			__/5		
Tot.	/12	__/15	__/52	__/48	__/20

General

1 An epidemiologist writes, in a grant application,:

$H_0 \quad \bar{y} = 10.0 \qquad H_{alt} \quad \bar{y} < 10.0$

Comment.

2 Separately for each of the following, say whether it is an example of a matched pairs design. Explain.

___ A teacher compares the pre-test and post-test scores of students.

___ A teacher compares the scores of students using a computer-based method of instruction with the scores of other students using a traditional method of instruction.

___ A teacher compares the scores of students in her class on a standardized test with the national average score.

___ A teacher calculates the average of scores of students on a pair of tests and wishes to see if this average is larger than 80%.

3 The water diet requires one to drink two cups of water every half hour from when one gets up until one goes to bed, but otherwise allows one to eat whatever one likes. Four adult volunteers agree to test the diet. They are weighed prior to beginning the diet and after six weeks on the diet. The weights (in pounds) are

Person	1	2	3	4
Weight before the diet	180	125	240	150
Weight after six weeks	170	130	215	152

Calculate the sums, W^+ and W^-, of the signed ranks.

4 A sociologist is studying the effect of having children within the first two years of marriage on the divorce rate. Using hospital birth records, she selects a random sample of 200 couples who had a child within the first two years of marriage. Following up on these couples, she finds that 80 couples are divorced within five years. To determine if having children within the first two years of marriage increases the divorce rate we should

___ test the hypotheses H_0: $p = 0.50$, H_a: $p \neq 0.50$.

___ test the hypotheses H_0: $p = 0.50$, H_a: $p > 0.50$.

___ test the hypotheses H_0: $p = 0.40$, H_a: $p > 0.40$.

___ do none of the above

Explain your answer.

Why is osteoarthritis of the hip more common on the right?

Newton J et al., The Lancet Vol 341 pages 1207, 7 May 29, 1994.

Note: the authors reported the relative frequencies in terms of the right/left ratio, whereas you might have been more comfortable with the proportion $\hat{p}$ of right sided THR's. This use of the ratio of right to non-right ($\hat{p}_{right}/\{1-\hat{p}_{right}\}$) is similar to the use by demographers of the male/female (i.e. male/non-male) ratio: e.g., if the proportions of male and female births are 0.51 and 0.49 respectively, demographers calculate the "sex ratio" as 0.51/0.49 = 1.04, i.e., 1.04 males for every female.

For parts 1 to 4 below, restrict your attention to the data from Oxford and Avon (first row of table).

1 State the implied null and the (1-sided) alternative hypothesis[1] in terms of

(a) the proportion p

(b) the equivalent parameter, the ratio p/(1 – p)

[p/(1 – p) is known to epidemiologists as the odds]

and calculate a statistic, and a p-value, to test it.

Do your p-value and conclusion match the "The frequency of unilateral primary hip replacements was significantly higher on the right than the left" and the "p < 0.01" reported by the authors?

[1] Although it is not appropriate to make a "right-sided" hypothesis after seeing the data, suppose that someone had—ahead of time—predicted that we would 'wear out' the right hip first.

Osteoarthritis...

2 Would the p-value obtained from a X^2 test be appropriate for evaluating your 1-sided hypothesis above? Explain.

3 How does one calculate a 95% CI for p
[calculation not necessary, but answer, to 2 dp, is 0.61 to 0.71]

From these limits, calculate a 95% CI for the ratio p/(1 – p).

[HINT: to obtain the CI for the ratio, evaluate the function p/(1 – p) at p = p_{lower} and p = p_{upper}. You did something similar when calculating a CI for "the number required to treat" in a previous exercise]

4 Refer again to H_0 in question 1. From just the reported CI of 1.52 to 2.48 for the ratio, and without any further calculations, what can you say about the 1-sided p-value? the 2-sided p-value? .

5 Suppose that instead of data reported in the table, all you were told was that "all four ratios were greater than unity", or (equivalently) that "all four $\hat{p}$'s were greater than 0.5".

What statistical model/distribution/table could you use to measure the strength of this evidence against the null hypothesis.

Would you be convinced if there had been 10 data sources and in 8 of the 10, the ratios were greater than unity? Why?
[formal calculations not required, but carefully explain your reasoning]

Mega-dose vitamin C in treatment of the common cold: a randomised controlled trial" Audera C et al., MJA Vol 175 pages 359-362, 1 Oct, 2001. eMJA: http://www.mja.com.au

Introduction [immediately following the abstract]

1 Turn the question "Would vitamin C, ..." in the last sentence into formal statistical hypotheses, written with symbols.

Statistical analysis

2 Assume that the sample size calculations were based on a comparison of 2 groups, with no correction for multiple comparisons. Besides this assumption, and whether they planned 1- or 2-sided tests [not stated, but assume 2-sided], what other one piece of information, not explicitly stated in the methods, would you need before you could reproduce their sample size calculations?

3 What formula or table from the course is appropriate for calculating the "desired sample size" here? *[calculation not required. A reference to the table of formula, with the appropriate inputs identified, is sufficient].*

4 From the standard deviation of 4 days, and the mean of 7, what shape did the authors anticipate for the distribution of durations?

5 Does this pattern have an impact on the validity of the statistical tests they planned to use? Explain.

Results/study population

6 If you compared 2 specific (pre-selected) treatment groups on the basis of 20 separate and unrelated baseline characteristics, such as the last digit of their phone number, which day of the month they were born, the temperature the day they were born, the number of siblings they had, how many courses in statistics they had had, etc., how many of the 20 comparisons might you expect to be significantly different at the $P < 0.05$ level?

vitamin C: Results/study population ...

7 What is the probability that at least one comparison would be significant? [Hint: answer can be found in table C of M&M]

8 If you now compared every possible combination of one or more of these 4 intervention groups in the study with one of more of the remaining groups(s), what is the impact of this on the chances of finding at least one significant difference?

Results/Cold duration and severity

9 [along the same lines...] What is the impact of making outcome comparisons between groups on the basis of individual symptoms and timepoints (1st paragraph) and groups that have the lowest medians (2nd paragraph)?

10 The authors say ["Actual power of study" paragraph] that, overall, the SD for duration was 6.6 days. Does that fit with the widths of the reported CI's [in box 2] for the days of symptoms, where, give or take a little, each sample size is approximately 46, and each CIhas a margin of error of approximately ± 2 days?

Box plots / Cold duration and severity

11 Because the cumulative severity scores are somewhat skewed, with a long upper tail, someone suggests that the t-tests are not valid and that one must use non-parametric tests instead.

a This is somewhat of an over-reaction, Explain why.

b What non-parametric test is appropriate here for comparing outcomes

between two specific groups?

among all 4 groups at once?

vitamin C: Doctor visits/Other medications taken [Box 2]

12 The percentages for doctor visits are accompanied by asymmetric CI's. Why is this? The authors probably got them from a table, or from a spreadsheet. But how did those who made the table or spreadsheet derive/program them?

The percentages who took other medications are accompanied by more symmetric CI's. Why is this? How can one calculate symmetric CI's for such percentages? [*don't calculate .. give a reference]*

13 What test(s) would one use to formally compare the percentages [give references]

• between 2 groups

• among all 4 groups at once?

Blinding / Cold duration and severity

14 The authors had only 31 recorded guesses by which to judge their success in maintaining blinding [3rd paragraph] and even then the way they report them is a bit confusing.

Assume that the 47+50+45 = 142 subjects who got 1g or 3g knew that there were to be 4 equal sized treatment groups [e.g. they were shown the text in the Interventions section]. Also assume that the question was posed as a forced binary choice:

"Do you think you have taken ≥1 g?" __ *Yes* __ *No*

How many of the 142, just by guessing alone, would you expect would correctly guess their assignment?

Say 110/142 guessed correctly. Without any formal calculations, what would you conclude?

What if the number was 125/142?

Sketch what formal calculations you might do based on the 125/142

vitamin C: Actual power of the study / use of CI's instead

PREAMBLE: *The data have spoken! Assuming no biases, one could ask: how 'definitively negative' is this finding? Instead of recalculating the detectable difference, with the same power and sample size, to be 40%, why not calculate the differences in average duration that are "ruled in" and "ruled out" by the observed data (i.e.. why not calculate say a 95% confidence interval for the true difference)*

To answer this next question (15), concentrate on 2 specific groups: the 0.03g versus 3g comparison.

15 The observed means are 10.4 days (SEM approx. 1 day) and 8.5 (SEM again approx. 1 day) . Of interest is the true reduction in the mean number of days of symptoms by taking 3 rather than 0.03g of vitamin C.

Under the "best case-scenario", and from a rough calculation, what is the greatest reduction that is compatible with the observed data?

Re-enrollments [See Methods/Participants and Abstract]

16 Participants were eligible to re-enroll in the study. Assuming nobody enrolled more than 2 times, the 184 cold episodes in the 149 people represent 184 – 149 = 35 persons who had 2 episodes, and therefore 149 – 35 = 114 who had 1 episode.

Technically speaking, this 'recyling' of participants affects the validity of the statistical procedures the authors used. Explain.

Some textbooks use strong warnings about this violation of the assumptions of the commonly used inferential procedures. As a result, many authors would, in this example, be afraid to use the full 184 observations in the analysis. Explain how - at the design, and even at the analysis, stage -- this re-enrollment was actually an opportunity to make the study more statistically efficient.

Is Age Associated with Size of Adult Extrahepatic Bile Duct: Sonographic Study

Horrow MM, et al. Radiology Volume 221 pages 411-414, November 2001.

INTRODUCTION

1 There are a number of possible meanings of the term 'normal' in this first paragraph ('center' of the distribution? or some upper percentile of this distribution?).

Put this issue of terminology aside for now and consider the 'simple rule of thumb', namely "a 4 mm mean duct diameter at age 40, a 5 mm mean duct diameter at age 50, and so on".

Fill in the blanks to paraphrase this rule for 'normal' as

'normal' for given age = _______ + _________ × age

MATERIALS AND METHODS

2 "Statistical analysis was used to test the hypothesis that ..." [beginning of last paragraph]

Write out the two 'competing' hypotheses; use symbols, and take care to choose correctly between Roman and Greek, and hats or no hats!

RESULTS

3 Comment on how appropriate the shape of the age distribution was for the objectives of this study. Had you had a say in the patient selection, would you have done it differently?

Note for the analyses discussed below, that although the authors give the distribution of all 1540 measurements (mean 3.5mm, SD 1.2mm), they wisely used only 1 number per person (either the mean of the 6, or the proximal AP measurement) as the "y" in each of their separate analyses.

4 Use the fact that the mean of the 258 y values is just about 3.5 mm, that the average age is 55, and the reported slope of 0.000578 mm/year to plot the fitted equation on the graph.

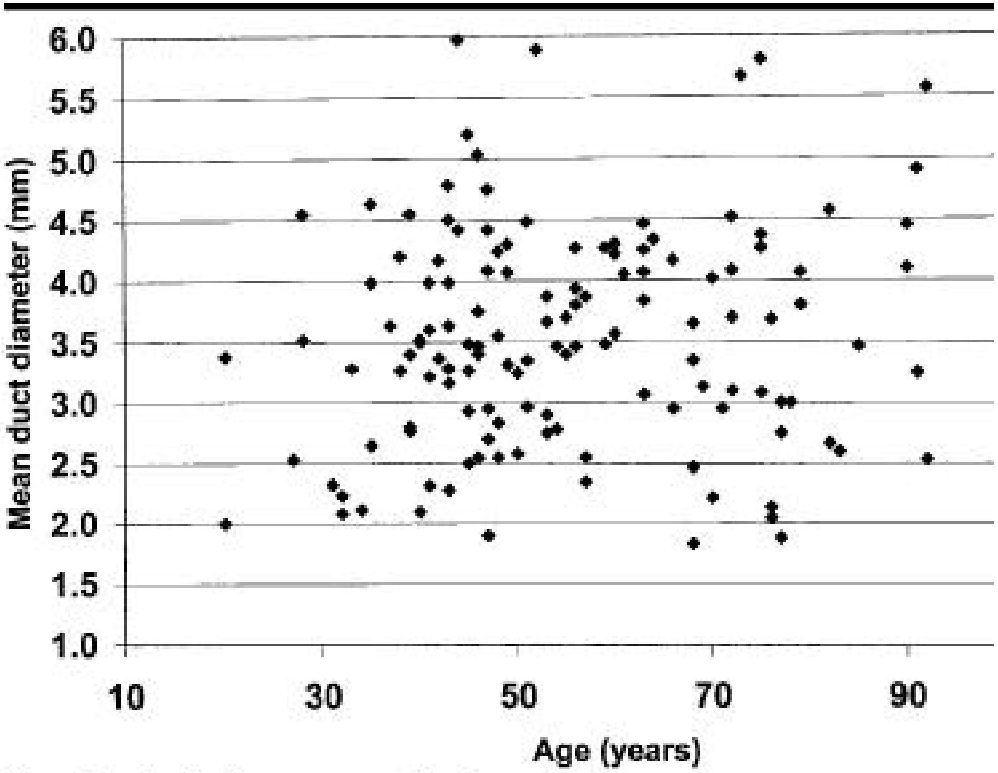

Figure 2. Graph depicts average duct diameter versus age.

Extrahepatic Bile Duct and age...

5 Remembering that they set up their hypothesis as 1-sided, verify from the reported slope and its SE that it is indeed significantly lower than 0.1mm/year (p < 0.001).

6 "Moreover, a 95% CI contains zero, indicating failure to provide evidence of an association of duct diameter with age"

This is tantamount to saying that the slope is not significantly different from 0 at the α =0.05 level.

Explain how the authors could have made better use of the CI than simply using it as a way to test if a slope is zero (hint: make use of all those zeroes in the CI!)

How about all those zeroes? (preface to question 7):

The reported SE for the slope is *very* small. It could be a type III error! In my notes on simple linear regression, I re-expressed M&M's formula for the SE of the estimated slope as (approximately)

$$\text{SE(slope)} \approx \frac{\text{RMSE}}{\sqrt{n} \times \text{SD(x)}} \ .$$

Now, $\sqrt{n}$ is approximately 16, and by coincidence, SD[age] is also 16, so

$$\text{SE[slope]} \approx \text{RMSE} / (16 \times 16) = \text{RMSE} / 256.$$

So all we need to recalculate the SE[slope] is the RMSE. We can estimate the RMSE visually from the graph: since the slope of the fitted line is itself unbelievably close to zero, the visual estimation is easier than usual.

7 Skip the summing of squared residuals, the dividing by n-2, and the taking of a square root. Instead, estimate the RMSE by eye as the "average" or 'typical" absolute deviation from the (here, flat) line

$$\hat{y} = 3.5 + 0 \times \text{age} = 3.5 \text{ mm.}$$

_____ is my visual estimate of the RMSE (mm)

Divide your estimated RMSE by 256 to get your best estimate of SE[slope], and compare it with the reported SE of 0.000334

<u>Comment</u>: maybe with the conversions between 1 mm per decade, and 0.1 mm per year, somehow an extra 10 got into the SE[slope]?

And if it did, could it be that the reported slope is also smaller than it should be? This is a bit like Mendel's monks, who got a chi-square goodness of fit statistic so close to perfect (0) that people became suspicious.

If indeed there is a type III error in the SE[slope] (and maybe even one in the slope itself?) is it serious? In the note which the authors must now send to the journal, what should they say about how it affects their conclusions?

Extrahepatic Bile Duct and age...

8 On the graph, using separate pen colours, or broken and unbroken lines, draw your best guess for

(a) the 95% confidence limits for the mean y
(b) the 95% prediction limits for individual y's.

9 As best you can, from what the authors reported, and from what you can see on Figure 2, fill in the blanks in the following table

```
PROC REG data=bileduct;
  MODEL  mean_duct_diameter = age ;
```

```
Root MSE       _______     R-square   ________

Dep Mean       _______

            Parameter  Standard   T for H0:
Variable DF  Estimate   Error     Parameter=0 Prob>|T|

INTERCEP  1 __________  _________   _______    _______

AGE       1 __________  _________   _______    _______

                 Analysis of Variance
                 Sum of       Mean
Source     DF    Squares     Square   F Value  Prob>F

Model      1     v. small   v. small  _______   _____

Error     ___    _______    ________

C Total   257    _______
```

10 In light of the findings of this paper, rewrite your version of what the 'rule of thumb' should be.

Birthweight, early environment, and genetics: a study of twins discordant for acute myocardial infarction.

Hubinette A, et al. The Lancet Vol 357 pages 1997-2001 June 23, 2001.

Table 1

1 Birthweight comparison

Explain why you do not have enough information in the table to verify the p-value.

set up the formula to carry out an un-paired test.

Table 3

2 Suppose you wanted to test the association between AMI and being the first-born twin of the pair. [Hypothetical: data on this were not reported]]

Show how to set up the data and how to analyze the results

3 Suppose you wanted to test the association between AMI and the Apgar Score[2] at birth. Say you are reluctant to use it as a quantitative variable, or to calculate mean scores.

Propose a statistical test, and give a reference [from the course] to a worked example of this test. Make sure your research assistant could follow the steps.

Introduction

4 The authors, in the third paragraph of the Introduction, state that "even within pairs of twins of the same sex, there are commonly substantial differences in birthweight"

Why is this important in this study?

[2] APGAR Scoring for Newborns: A score is given for each of 5 signs [Activity (Muscle Tone); Pulse; Grimace (Reflex Irritability); Appearance (Skin Color); Respiration] at one minute and five minutes after the birth. If there are problems with the baby an additional score is given at 10 minutes. A score of 7-10 is considered normal, while 4-7 might require some resuscitative measures, and a baby with Apgar scores of 3 and below requires immediate resuscitation.

AMI/twins...

5 One could paraphrase the statement in **4**. as "Birthweights of twins are far from perfectly correlated".

Say that the observed correlation is of the order of 0.6, and that associated p-value is < 0.001. What does this p-value mean?

6 Freedman, on page A-7 of his text Statistics, says that in twin studies, the convention is to plot each twin pair twice: once as (x,y), and once as (y,x).

Suppose you did this with 132 twin pairs (i.e., you created 264 datapoints), but forgot to take this 'sample size inflation' or 'data cloning' into account when calculating a CI based on the observed correlation. In other words, you based the CI on the 'n' of 264. How much narrower is your CI than it should be?