Take Home Exam Course EPIB-694, June 2004

Instructions

- Completed exam is due by 5pm June 25, 2004.
- Teams may be of size 1, 2 or 3
- See content of Exam for 1, Exam for 2, or Exam for 3.
- Until June 25, no discussion or communication concerning the exam questions/answers with any person outside of your team (other than JH).
- One set of answers per team.
- Set of answers to be accompanied by a joint statement (see last page) as to contributions of each member and a declaration that there was no outside help [declaration required from solo efforts too!],

  A team member who is uncomfortable with what was described in this joint statement may send JH a separate, independent statement as to what he/she believes his/her and others' contributions were. JH will keep this confidential.

Content of exam, as a function of team size

	Exam for 1	Exam for 2	Exam for 3
Do dogs resemble their owners? questions a to f	X	X	X
Do dogs resemble their owners? questions g to i		X	X
Do dogs resemble their owners? questions k to n			X
Why is osteoarthritis of the hip more common on the right?	X	X	X
OSIRIS trial, questions 1, 2, 5, 6, 7, 8, 9	X	X	X
OSIRIS trial, question 10		X	X
OSIRIS trial, question 11			X
Helicobacter pylori infection and gastric cancer, questions 1, 2, 5 to 10	X	X	X
Helicobacter pylori infection and gastric cancer, questions 11 to 14		X	X
Helicobacter pylori infection and gastric cancer, questions 15 to 17			X
Differences in proximal femur bone density over two centuries	X	X	X
Optional ...			
Why do old men have big ears? questions 3 and 5	X	X	X
Why do old men have big ears? questions 2 and 4		X	X
Why do old men have big ears? question 6			X

Do dogs resemble their owners?

a *If a judge simply guessed (or tossed a coin) to decide which of the two dogs belonged to the owner, what is the probability that the judge's guess would be correct?*

b *If 28 judges guessed, what is the probability that (i) a majority i.e., more than 14 (what the authors call a 'match') (ii) exactly 14 (what the authors call a 'tie') and (iii) fewer than 14 (what the authors called a 'miss') would match the owner and the dog?* The Excel spreadsheet "Binomial Distributions (how shapes varies with n and p) in the Resources for Chapter 5 of course 607 can help you here.

c *If 28 judges guessed about 20 different owners (and their non-purebred dogs), for how many of the 20 would you expect there to be a match? a tie? a miss?*

d *How do these expected numbers compare with the numbers observed in the study (first paragraph of Results)?*

e *If, instead of the 3 categories the authors used, you used a simple dichotomy ">14" versus "14 or fewer" (i.e. a 'match' versus 'ties or miss', and if indeed judges were simply guessing, what is the probability of observing (i) 1 match in 1 owner (and its non-purebred) dog (ii) 7 or more matches in all 20?*

f *Repeat above calculations, but for the 25 purebred dogs.*

If instead of transforming the number, out of the 28 judgments, that were correct, into 2 or 3 categories, suppose you used the number/28 'as is' i.e., as a *number* between 0 and 28 (as in the raw data Table).

g *Under the null hypothesis that the judges were simply guessing, what is (i) the mean (i.e. expected) value (ii) variance and (iii) standard deviation of this number?*

h *What does the (alternative) hypothesis that the authors wished to test say about the expected value?*

i *Across the 25 purebreds, what is the average number of judges who correctly matched owner and dog?*

j *Using **g-i,** calculate a test statistic, and its associated (one-sided) p-value. Comment.*

In the majority of applications involving tests of means, one must estimate the variance or standard deviation from the data, and use the (wider) t-distribution to account for the extra uncertainty; here, in this example, under the null hypothesis, you know the variance. Thus , if under the null, the observations would have a Gaussian distribution, one could use the Z-distribution as the reference distribution.

k *In this situation, under the null hypothesis, is it reasonable to assume that the numbers would have a close-to-Gaussian distribution around the mean of 14?*

Imagine that the investigators had designed a more difficult test, where instead of *one* other dog, they had six other dogs.

l *Under this scheme, what would the expected (mean) value and the standard deviation of the number of judges who picked out the correct dog?*

m *Under this scheme, Why would the number not have a close-to-Gaussian distribution around this mean?*

n *Would one still be justified in using the Z-test to test whether across the 25 purebreds, the average number of judges who got it right was significantly higher than expected under the null? What if there were only 5 purebreds? what if there were 100?*

o Suppose you boss/chief (or the editor/referee for the journal) had never heard of the Central Limit theorem, is not convinced by your answers, and suggests that you perform a non-parametric test of the same null hypothesis for the 25 purebred dogs.

- i *List 2 such tests* (we have already used a variation on one of these, without giving it a formal name), *and indicate which should be the stronger (more powerful/sensitive) of the 2*
- ii *State the null (and alternative) hypothesis they test.*
- iii *Carry out the two tests and comment on the findings.*

p The authors compared the classifications of the 20 non-purebreds with those of the 25 purebreds, since their theory predicted that the accuracy with the purebreds should be better.

- i *Use the 3 different data-reduction methods (trichotomy, dichotomy, raw number correct) to compare the accuracy in the 20 versus the 25.*
- ii *Give two reasons why, for this type of situation, their chi-square methods should not be as sensitive as those based on the non-categorized numbers (Hint: one has to do with the 'granularity' of the data, the other with which tests do/do not take account of the directionality in the alternative hypothesis)*

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?

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]

[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.

Early versus delayed surfactant (OSIRIS Study)

Statistical methods

1 *Write, in symbols, the formula you would use to calculate the required sample size for the "early vs delayed selective" portion of the study.*

Briefly explain each symbol, and say what value you would use for it.

Why is the calculated sample size (2000) so large?

2 *Explain what is meant by the word "power" in the phrase "on the assumption of 80% power and ..." [line 12]*

Table I

3 *Many authors (and even reviewers) mix up SD and SEM. How can you be sure that the 2.31 weeks of gestational age is not an SEM? [3rd row, 1st column]*

4 *Why do you think the authors used mean [SD] to describe birthweight but Median {IQR} to describe age at entry?*

Table II

5 *If you wanted to compare the average number of doses administered to the early vs delayed selective groups, what statistical test would you use?*

6 *Do you think that any of the requirements for the validity of this test are seriously violated in this example? Why? / Why not?*

Table III

7 *Write out the formula used to get the 95% CI of -9.9 to -2.7 [last row, 4th column]; use numbers in the formula but do not complete the calculation.*

8 *List the steps followed to obtain the p=0.057 [2nd row, 5th column], imagining that you were explaining them to a research assistant; do not complete the calculations.*

9 *Use the results in columns 4 and 5 to illustrate how one can perform tests of significance directly from CI's without additional calculations.*

Dosing Comparison

10 ***"the outcome was similar in the two groups in respect of all principal measures of outcome... and in respect of all secondary measures" [1st sentence, 2nd paragraph]***

"the trial provides no evidence that an "up-to-4-doses " regimen is superior to a regimen of 2 doses " [last sentence of Abstract]

You are the neonatology resident; the head of neonatology is a stubborn supporter of the "up-to-4- doses" regimen and when you mention this study to him, he throws words like "inadequate power" and "type II error" at you. Briefly, what do you say [statistically speaking] to him to try to convert him?

Overall

11 This report uses both ratios and absolute differences when comparing outcomes.

Which one do you prefer for which purposes? Why?

An international association between Helicobacter pylori infection and gastric cancer"
EUROGAST Study Group; The Lancet Vol 341 pages 1359-1362 May 29, 1993.

Subjects and Methods

1 Three populations [from US and Japan] were added later to extend the range of gastric cancer incidence. (2nd and 3rd sentences of 1st paragraph)

In this study, the value of "Y" (incidence) was known before the value of X (seroprevalence) could be measured. If one could choose populations on the basis of their X (rather than their Y) values, why would it help to choose ones which "extend" the range of X?

2 "We aimed to recruit 50 males and 50 females in each of the two age groups... (same paragraph)

How precisely can one measure the seroprevalence in these sex-age groups with these sample sizes?

As we will see later, it is reasonable to pool the male and female samples in order to better estimate the common (unisex) seroprevalence in a centre.

What numbers would be required per centre so that, for most centres, we could expect the estimates of their seroprevalences to be within 5 percentage points of their true seroprevalences? [focus on one age group; however as we will see later, the authors averaged the prevalences across age groups]

3 "The sensitivity and specificity of this test was 96% and 93% respectively" (end of first paragraph)

Express these two percentages as conditional probabilities.

4 ... the line which best fitted the data...

Explain to your friend, who studies history, the criterion by which one determines the line which "best" fits the data.

5 Cancer rates were log-transformed...

Why do you think the authors did this?

6 The seroprevalence for each center was calculated as the average of the two prevalences...

Compared with the precision of each separate seroprevalence estimate, how much more precise is the average of the two seroprevalence estimates?

Results

7 "There was therefore a nine-fold range in seroprevalence in the younger age group" (middle of 2nd paragraph). In young males, for example, the observed *H Pylori* seroprevalence varied from 8% to 70% across centres. Assuming each % is based on approximately 50 subjects, we can test if this centre-to-centre variation is more than just (random) sampling variability. A X^2 test, with (16)(1)=16 d.f. applied to a 17 x 2 table of the frequencies of seropositive and seronegative subjects yields a test statistic of approximately 140, which is "off the map" of the reference X^2_{16} distribution[2].

Why is it important first to establish that the observed variation in seroprevalence is significant (i.e., "real / non-zero") ?

8 "Within each of the individual populations the prevalence was higher in the older group than the younger one" (next sentence). Your chief uses p-values the same way a drunk uses the lamppost—more for support than illumination!. Even though this pattern makes good biological sense, at the journal club he still needs a p-value to be convinced that it is more than just coincidence or a "fluke". You don't have a calculator or set of statistical tables handy, but you want to impress him by coming up with a p-value before the journal club ends.

---

[2]This is an example where an omnibus test of

H_0: $\pi_1 = \pi_2 = \ldots = \pi_{17}$

makes sense, since we have no obvious alternative hypothesis other than the non-specific

H_{alt}: there is **some** variation among the17 centres (π's)

What fast test of significance might you do to humour him? Do it on the back of an envelope and explain to him the logic behind it. He has never had a course in statistics but always asks for a p-value, especially when there are others around, just to impress them.

9 "But there was a strong correlation between the prevalence at 25-34 years and that at 55-64 years, r = 0.88, both sexes combined" (second half of sentence). Again, your boss asks for a p-value. You get a little annoyed at this point. You are tempted to tell him about the message "God is the answer!" that you saw written on a bathroom wall, below which someone (probably an epidemiologist!) had written "Yes, but what is the question?". But you know better than to embarrass him, and so you think of a less hostile answer.

What is your answer?

10 "There was no appreciable difference between the prevalence in males and females [36% and 34%, respectively]" (next sentence). You calmly explain to your boss that even if the 2% difference is statistically significant, it is not a meaningful difference, and that this is why the authors said it was not an "appreciable" difference. He agrees, but now says, "yes I know, but I still want you to explain what statistical tests you are learning over in that epidemiology and biostatistics department that would be appropriate here"

What test would you do, and how would you do it? You don't have to do the test; you can just give a reference.

Then explain to your boss why a confidence interval might be more meaningful here than a statistical test.

11 The regression analysis was done with log-transformed rates, with logs to the base *e*, the natural logarithm, i.e., *ln* (rates). Thus the rates are displayed on a log scale, but on what looks like to the base 10. To line the datapoints up with the base *e* that was actually used in the regression analysis, JH has added in the *ln* scale on the vertical axis of each graph in the figure. Thus the point on the axis marked 0.1 corresponds to *ln* (0.1) = –2.3, the rate of 1 to *ln* (1) = 0, the rate of 10 to *ln* (10) = +2.3, etc. Thus, for example, (to 1 decimal place)

ln (mortality rate, Florence males) = *ln* (3.0) = 1.1,

ln (mortality rate, Minneapolis St. Paul males) = *ln* (0.6) = –0.5.

[The scatterplot in the top left panel should match the scatterplot of *ln* rate versus prevalence in the computer printout below]

By hand, using the ln scale, measure the slopes of the 4 fitted lines, and see how close you get to the regression coefficients given in the four panels. Comment!

12 "For each sex, there was a significant relation between seroprevalence and log-transformed mortality and incidence rates" (first sentence, third paragraph)

What steps does the statistical software go through to determine the p-values shown in the figure?

13 In the combined model, the coefficient was 1.79 for mortality—i.e., a 10% increase in infection prevalence was associated with approximately an 18% increase in log (actually *ln*) cancer mortality. (next sentence)

Explain how they arrive at the 18%.

14 Although there was a clear association..., there was also considerable scatter (5th paragraph)

What number is usually used to measure the scatter? What is this number in the printout below?

15 From this printout, extract

i *the average (mean) ln mortality rate*

ii *the variance of the 17 ln mortality rates [by this, I mean the variance about the mean, defined in M&M Chapter 1, not the variance about the regression line that is the focus of Chapter 10]*

iii *the variance about the regression line*

iv *the p-value from a test of whether the correlation between ln rates and H Pylori seroprevalences is zero.*

v *the fitted (or predicted) ln mortality rate in populations that have no H Pylori infection*

vi *take the antilog of this number, i.e., exp[this number], to get the fitted (predicted) rate of gastric cancer mortality for populations that have no H Pylori infection*

vii *the predicted ln mortality rate for populations with 100% H Pylori infection*

viii *exp[this number], ie. the fitted or predicted risk of gastric cancer mortality in populations with 100% H Pylori infection*

ix *the ratio of viii to vi.*

16 "After accounting for sex, the proportion of the variance in the log-transformed cancer rates explained by *H Pylori* positivity was 18.3% for mortality". (last sentence). It is not clear how exactly the authors "accounted for sex". But one straightforward way to do so is to examine the relationship within each sex.

Within males, how much of the variance in log mortality rates is explained by H Pylori positivity (see printout)?

17 The authors are quite open about the limits of correlation studies, and their "implicit assumption" at the bottom of the second column of the Discussion. One factor, which they did not discuss, is the fact that the seroprevalence for each centre is estimated from a fairly small sample, and thus subject to sampling error.

What effect does this have on the observed relationship of seroprevalence and mortality? In other words, imagine it were possible to measure seroprevalence on <u>everybody</u> in these populations. If it were, would you expect that the slopes would be (i) steeper (ii) shallower (iii) about the same as those obtained with "error-containing" estimates of seroprevalence?

Analyses of gastric cancer mortality rates for males

`pr_m:` prevalence(proportion) of *H Pylori* in males
`lnMort_m:` *ln* mortality in males

```
data sasuser.h_pylori;

INPUT Center $ Mort_m    Mort_f   Inc_m    Inc_f
  Pr2534_m  Pr2534_f Pr5564_m  Pr5564_f Total_n;

pr_m = (Pr2534_m + Pr5564_m)/200;
lnMort_m = log(Mort_m);

LINES;
AL  1.6 0.7 1.6 0.7 42  44  49  69  200
GH  1.1 0.7 1.2 0.6 20  17  60  47  208
...
...
MS  0.6 0.2 0.9 0.3 13  16  36  32  198
;
RUN;
```

```
PROC PLOT data=sasuser.h_pylori;
  PLOT lnMort_m * pr_m;
RUN;
```

```
Plot of LNMORT_M*PR_M.  Legend: A = 1 obs, B = 2 obs, etc.

LNMORT_M |
       2 +
         |
         |                                          A
         |
         |                                 A   A
         |              A                       A
         |       A
       1 +                           A
         |            A    AA          A
         |
         |
         |                   A
         |        B
         |               A
       0 +
         |
         |
         |
         |     A                            A
         |
         |
      -1 +
         |
         ---+------------+------------+------------+--
           0.2          0.4          0.6          0.8
                             PR_M
```

```
PROC REG data=sasuser.h_pylori;
  MODEL  lnMort_m = pr_m ;RUN;
```

(**SAS**) Dependent Variable: LNMORT_M

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	1.60445	1.60445	4.788	0.0449
Error	15	5.02620	0.33508		
C Total	16	6.63065			

(**SAS**)

Root MSE	0.57886	R-square	0.2420
Dep Mean	0.74167	Adj R-sq	0.1914
C.V.	78.04809		

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.12	0.42	-0.279	0.7844
PR_M	1	1.75	0.80	2.188	0.0449

SUMMARY OUTPUT (Excel)

Regression Statistics	
Multiple R	0.49
R Square	0.24
Adjusted R Square	0.19
Standard Error	0.58
Observations	17

ANOVA

	df	***SS***	***MS***	***F***	***Significance F***
Regression	1	1.6044	1.6044	4.7883	0.0449
Residual	15	5.0262	0.3351		
Total	16	6.6306			

	Coefficients	***Standard Error***	***t Stat***	***P-value***	***Lower 95%***	***Upper 95%***
Intercept	-0.12	0.42	-0.28	0.78	-1.00	0.77
pr_m	1.75	0.80	2.19	0.04	0.05	3.46

Differences in bone density over 2 centuries

a "The precision of measurement" was 1.24% [3rd paragraph of Subjects and Methods]? It is not clear if the 1.24% is an average (or other summary) of three SD's or 3 CV's (Coefficient of Variation). The latter is more general, since it is independent of units. Suppose that for one of the three femora, the five measurements at one site were 0.84, 0.86, 0.86, 0.88, 0.81. Calculate the coefficient of variation (CV) as CV = 100 x SD[5 measurements] / mean[5 measurements].

b What does the "SEE (standard error of the estimate)" measure? Explain how it is calculated. Under what other name is it found in the output of other statistical packages?

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

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

(i) Superimpose onto Fig 2 the fitted regression line for the ancient femora, femoral neck, premenopausal women.

(ii) One way to test for a non-zero slope is via the statistic: slope/SE[slope], vs. t-distrn.. What information is needed to calculate the SE of the slope? What 3 factors influence the magnitude of the SE of the slope? [the alternative form for the SE of a slope, in my notes for M&M Ch 2/10 might help]

(iii) Use the information in Tables I and III to reconstruct the SE of the slope [for the ancient femora, femoral neck, premenopausal women] and calculate the test statistic. Interpret the result. [Another way is to test for a non-zero correlation -- since in simple linear regression there is a 1:1 relation between the slope and correlation]

(iv) Calculate a 95% CI for the slope. In view of this, can we take the statement about "no significant loss" above as a definitive statement about the absence of premenopausal loss?

e "in striking contrast to modern women" refers to the slope of 0.197 vs. that of –0.658 (neck) and –0.162 vs. –0.921 (triangle). Calculate a SE for the –0.658. Use it and the one for the 0.197 to verify that this "striking difference" was indeed statistically significant. [#].
If you were an editor, and -- for space reasons -- it was a choice in Table III between showing the column of "SEE's" and showing "SE's" for selected slopes, which would you choose? Why?

f The answer from the test involving r=0.424 was " * p<0.005"; upon what null hypothesis is the p value calculated?

$t = (b_1 - b_2)/SE[b_1 - b_2] = (b_1 - b_2)/Sqrt[\ SE^2[b_1]+SE^2[b_2]\ ]$

Why do old men have big ears?

James A Heathcote, British Medical Journal, December 1995, page 1668
[the Christmas Edition of BMJ is usually fun to read, even if you are not that fond of British humour .. JH et al. have a piece in the 2003 Christmas edition]

1 Unlike Epi-Info, many statistical packages do not return the 95% CI for B; instead, they report b and SE(b).

How does one go from b and SE(b) to the CI for B?

2 From the reported 95% CI for B, you can determine that the coefficient b is statistically significantly different from B=0 ($p < 0.05$ two sided).

But -- just from the CI-- can you calculate the actual p-value? If so, how?

3 The estimated mean ear length for patients of approximately 60 years is $55.9 + 0.22 \times 60 = 69.1$ mm. By substituting the lower and upper limits of the 95% CI for coefficient B into the equation $55.9 + B \times 60$, we obtain the limits

$$55.9 + 0.17 \times 60 = 66.1 \text{ mm}$$

and

$$55.9 + 0.27 \times 60 = 72.1 \text{ mm}$$

Compare this interval with the observed range of ear lengths for patients of age approximately 60 years. How do you explain the discrepancy between the calculated interval and the observed range of ear sizes?

4 *Does the report give enough numerical details to allow you to mathematically project what the observed range should be? If yes, do so. If not, explain.*

5 "It seems therefore that as we get older our ears get bigger"
[end of the Methods and Results section]

Given the data and the findings, is this inference justified? Explain.

6 *[challenging!] From the summaries given, and from assumed values when essential summary values are not reported, reconstruct the numerical output that would be produced by a regression procedure such as in SAS or Excel (for format, see examples in textbook pp 669 and 685, or in gastric cancer above). Carefully document your calculations and reasoning, indicating which items were taken directly from the report, and which you had to estimate 'by eye'.*

Epidemiology & Biostatistics, McGill University **Take Home Exam Course EPIB-694, June 2004**

Authorship Responsibility [adapted from JAMA]

Each author must read and sign the statements on **Authorship Responsibility, Criteria, and Contributions.**

Each author should meet all criteria below and should indicate general and specific contributions …

A I certify that the answers represent the work of the team members, and that no outside help was received (check)	____	____	____
B I have given final approval of the submitted answers.	____	____	____
C I have participated sufficiently in the work to take responsibility for (check 1 of 2)	___ part of the content [indicate which part(s)] ___ the whole content.	___ part of the content [indicate which part(s)] ___ the whole content.	___ part of the content [indicate which part(s)] ___ the whole content.
Your Signature			
Date Signed			
Your name (print or type)			