Instructions

- Completed exam is due back one week after receiving it.
- Teams may be of size 1, 2 or 3
- See content of Exam for 1, Exam for 2, or Exam for 3.
- Until December 19, no discussion or communication concerning the exam questions/answers with any person outside of your team.
- 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
Adult height of short children	X	X	X
Response Times of the Emergency Response System	X	X	X
Helicobacter pylori infection and gastric cancer	X	X	X
OSIRIS trial, questions 1, 2, 3, 4, 7, 8, 9	X	X	X
OSIRIS trial, questions 5, 6		X	X
OSIRIS trial, questions 10 and 11			X
Why do old men have big ears? questions 1, 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
Birthweight, early environment, and genetics, questions 1 to 7	X	X	X
Birthweight, early environment, and genetics, questions 8 to 11			X

Predicting the adult height of short children"
Heitmann BL et al., BMJ Vol 308 pages 1207, 7 May 29, 1994.

Methods and results

1 "The median age at this examination was 19 (range 18-27) years" (end of second paragraph)

(i) *Explain in words what the median represents.*

(ii) What do you think the mean age was? the mode? Explain your reasoning.

2 "Boys with heights below the height closest to the third centile are denoted boys with short stature" (end of fourth paragraph)

What does the "third centile" represent?

3 "Mean height of boys aged 7 increased from 124 (SD 6) cm to 176 (SD 7) cm in adulthood" (beginning of next paragraph). From this we know that the average (mean) growth was 176 – 124 = 52 cm.

From the SD's given, can you quantify the (between-person) variability (SD) of the amounts that individuals grew in that time frame? Explain your answer.

Do you think the SD of the individual amounts grown could be as much as 10 cm?

4 "For each age from 7-13... below the mean," (first half of next sentence). Say, for the sake of this exercise, that 25 had short stature at age 7, and that their (mean) eventual adult height was 1.8 SD below the mean.

(i) Calculate a statistic to test the 'common assumption' that children who are short before puberty will also be short in adulthood, and determine the associated p-value. For now, ignore the fact that the 25 are also part of the 912, and that the mean in the 912 is itself subject to sampling variability.

(ii) As the obsessive type, you separate out the 25 from the 912, leaving (912 – 25) = 887 who were not considered short at age 7. You calculate the mean (SD) adult height in the 25 and separately in the 887. How do you proceed from there to test the significance of the difference in adult stature? Be specific as to the steps, but do not do any calculations. Do you think the answer will be very different from that obtained in (i)? Explain your reasoning.

5 "and 30-43% of these boys remained below the third centile in adulthood" (second half of same sentence). Again, say, for the sake of this exercise, that 25 had short stature at age 7, and that 10 of them remained below the third centile in adulthood.

(i) Using this simpler definition of "short in adulthood", compute a p-value to test the "short as a child, short as an adult" hypothesis. Do so both ways: using the conservative approach explained in 4(i) and using the "correct, but probably doesn't change the conclusion" approach followed in 4(ii).

(ii) Would Gaussian approximations help here? Explain.

6 "None of the boys with short stature had adult heights above average" (first half of next sentence). Again, this way of looking at it provides strong evidence for "the common assumption" [cf 4(i)]. Nevertheless, suppose someone pressed you to test the hypothesis formally using this observation. Say, again, that we divide the boys on the basis of stature at age 7, where we have 25 'short' and 887 'not'.

(i) State the null hypothesis. Under this null hypothesis, how many of the 25 would you expect to have adult heights above the 'average'? (state any additional assumptions you would have to make, but don't fuss about the fact that the 25 contribute to—and drag down—the 'average')

(ii) suggest a "simple to compute, and easy to explain to your boss" way to calculate a p-value under this null hypothesis.

7 Refer to the "all" curve in the Figure.

(i) explain carefully what this curve represents, and in particular what the "percentage" on the vertical axis represents.

(ii) Suppose you did not have access to the mean and SD reported in the text, and had to 'extract' them from this curve. Explain how to do so, stating any assumptions you would have to make.

8 The authors selected a one per cent sample from the 93800 records available.

Had they selected say 50%, or 100% of the records, do you think the SD of the adult heights would have been larger than the 7 cm observed in the study? Explain.

9 "All men may not have achieved final height when measured at the draft board. In 313 subjects measured 4-34 years later height increased 0.5 (1.4) cm;" (last two sentences of first paragraph of Comment)

(i) Suppose that a person's true height cannot really 'shrink' during the 'later years' in question. Sketch the distribution of these 313 changes, under two different scenarios

(a) height is measured perfectly, with no error;

(b) height is measured with error.

(ii) *Is the observed increase of 0.5 cm*

(a) statistically significant ?

(b) important?

(iii) State the null hypothesis in (ii a)

Response Times of the Emergency Response System
(André Lavoie, PhD thesis, McGill 1992)

[no separate document; relevant facts given below]

The following data are taken from Table 4.8: EMS response time.(approx. 1500 instances)

	Response time (in minutes)			
Ambulance	Mean	S.D.	Range	Median
dispatch time	3.7	6.3	0 to 104	2
arrival time	5.9	3.8	0 to 46	5
response time	9.6	7.4	0 to 109	8

dispatch time: time from call to dispatch
arrival time: time from dispatch to arrival on scene
response time: dispatch time + arrival time

1 *Comment on the shapes of the distributions. What other summaries, other than those reported, might you consider interesting?*

2 *From the numbers given, show that there is virtually no correlation between the 2 variables dispatch time and arrival time*

3 *Suppose you had the raw data, and wanted to compute a correlation between the two variables. If you wanted to avoid any extreme influence of extreme values, what statistic would you propose? Explain the reasons for your choice.*

4 *Suppose you wanted to compare, in a formal statistical test, these response times with the corresponding response times after reorganization of the Emergency Response System. Would the shapes of the distributions be of concern to you? Explain.*

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[1].

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

---

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

H_0: $\pi_1 = \pi_1 = \ldots = \pi_1$

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)

ends.

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 epidermiology 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?

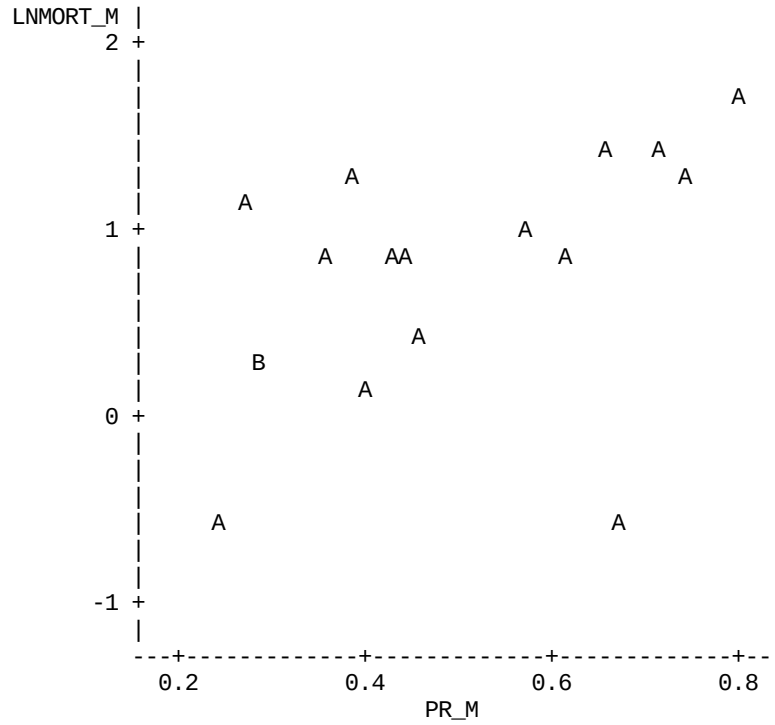
Analyses of gastric cancer mortality rates for males

`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.




```
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

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?

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]

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'.*

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.

Carry out an un-paired test instead.

Why do you get practically the same p-value with the unpaired test as the authors did with the paired test? Hint: read the description of the selection of external controls.

2 Marital status comparison (here, with just two categories, "test of homogeneity" is a test of two proportions)

JH suspects that in fact, the authors carried out an unpaired test. Do such a test and see on how close the p-value it yields comes to the reported p-value. Make sure you specify whether the alternative is 1- or 2-sided.

Although you don't have to, the authors probably used a X^2 test.

Explain why the X^2 test is sometimes described as a "1-tailed test, but for two-sided alternatives"

Table 2

3 Each of the odds ratios in Table 2 is 'adjusted' for the effects of the other variables in the table.

Is maternal age a significant independent predictor of AMI?

Why is there not a similar table of adjusted odds ratios after table 3?

Table 3

4 In the Summary, the authors interpret the "no significant differences in birthweight" and the p-value of 0.73 [Abstract, Findings] as a "lack of association" [Abstract, Interpretation].

How comfortable are you with their doing this?

How would a confidence interval accompanying the observed difference in birthweight have helped you to judge how 'definitive' the 'lack of association' is.

[$32,000 question] From what is given in the first row of Table 3, can you reconstruct the CI for the difference?

5 Suppose you wanted to test the association between AMI and being the first-born twin of the pair.

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

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

Propose a statistical test, show how to set up the data and how to carry out the test.

---

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.

7 As is explained in the middle of page 1998, the sample size was determined more by how many suitable birth records were available than by considerations of statistical power. Suppose that, nevertheless, one reviewer of the study plan wished to know how many twin pairs would be needed to have 80% power against an average difference of 100 g between affected and unaffected twins.

Identify and explain each of the components in the sample size calculation.

Describe how—ahead of time—you would obtain an estimate of the σ *that goes into this formula. Remember that this is for an analysis based on a paired test.*

Suppose you are unable to supply an estimate of σ. and instead planned the sample size as if you were going to use a test for two independent samples. But, suppose that at the time of analysis, you used a paired test, just like the authors did.

Would you still have 80% power against an average difference of 100g? Explain.

Introduction [$64,000 questions]

8 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?

9 One could paraphrase the statement in 8. as "Birthweights of twins are not perfectly correlated".

(i) Based on the matching criteria employed, how much correlation would you expect between the birthweights of the 118 pairs used in Table 1 (external controls)?

(ii) Using back-calculation from the summaries and the p-value given in the first row of Table 1, calculate the magnitude of the correlation.
Hint: Covariance$[Y_1,Y_2]$ = correlation$[Y_1,Y_2] \times$ SD$[Y_1] \times$ SD$[Y_2]$

(iii) Based on what you have seen of twin data in the course, or just based on your (un)educated guess, how much correlation would you expect between the birthweights of the 132 pairs used in Table 3 (co-twin controls)?

(iv) Again using back-calculation from the summaries and p-value given in the first row of Table 3, calculate the magnitude of the correlation.

(v) Would you expect the correlation to be higher for monozygotic than dizygotic twins? Why?

(vi) Repeat the calculations in (iv), but for monozygotic and dizygotic twins separately [note that 40+72=112, i.e.. for 132 - 112 = 20 twin pairs, zygosity was unknown]. Try to explain your findings.

10 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 (ie, you created 264 datapoints), but forgot to take this 'sample size inflation' 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?

11 Back to 'quantifying the substantial' [i.e. the magnitudes of the] within-twin-pair differences in birthweight.]

Calculate how much difference you could expect, and with what frequency. For example, in what percent of twin pairs might the difference exceed 200 g ? 400 g? State any assumptions or simplifications made.

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)			