Homegrown Exercises for Chapter 8 [Inference for proportions]

-1- Variability of, and trends in, proportions

The following data are the proportion of Canadian adults responding YES to the question "Have you yourself smoked any cigarettes in the past week?" in Gallup Polls for the years 1974 to 1985.

1974	52%
1975	47%
1976	---
1977	45%
1978	47%
1979	44%
1980	41%
1981	45%
1982	42%*
1983	41%
1984	39%
1985	39%

--- question not asked;
* question worded "occasionally or regularly"

Results are based on approximately 1050 personal in-home interviews each year with adults 18 years and over.

a Plot these percentages along with their 95% confidence intervals.

b Is there clear evidence that the trend is downward? To answer this, try to draw a straight line through all (or most of) the confidence intervals and ask can the straight line have a slope of zero i.e. be parallel to the horizontal axis. You might call this a "poor-person's test of trend".

-2- Dentifrices

In a study of the cariostatic properties of dentifrices, 423 children were issued with dentifrice A and 408 with dentifrice B. After 3 years, 163 children on A and 119 children on B had withdrawn from the trial. The authors suggest that the main reason for withdrawal from the trial was because the children disliked the taste of the dentifrices. Do these data indicate that one of the dentifrices is disliked more than the other?

-3- Sample size needed to asses risk of abortion after chorionic villus sampling

The following letter is by Holzgreve et al. to The Lancet (p. 223, January 26, 1985). They use symbols P_1 and P_2 in the same way we use the Greek (for "population") symbols "π_1" and "π_2". Also, they use the term 'rate' where we might use 'proportion' and they use it as a percentage i.e. their P_2=4.4% is our P_2=0.044. Note also that in the 1st sentence at the top of the page, they reverse the 2 subscripts. The correct subscripts are those used later on i.e. 1= ultrasonically normal pregnancies and 2=chorionic villous biopsy (cvb). Below, lower case p is used for a proportion observed in a sample.

> We agree with Dr Wilson and colleagues (Oct 20, p 920) that background rates of spontaneous abortion in ultrasonically normal pregnancies are an important requirement for evaluating the of chorionic villus sampling in the first trimester. For an unbiased assessment of the risk of spontaneous abortion with this new method of prenatal diagnosis, however, the rate of fetal losses should be compared with matched pregnancies without invasive procedures in a prospective, randomised trial.
>
> To be able to state with confidence that the fetal loss rate in a group of patients (P) after chorionic villus biopsy differs from that in a control group of ultrasonically normal pregnancies (P2) we have calculated the required sample size for the two populations, based on a probability of a type I error (a) of 1% and of a type II error (b) of 10%. The most recent international survey[2] revealed a spontaneous abortion rate of about 4.4% after chorionic villus sampling, and this was the figure we used for the rate in P2 when calculating sample sizes by the Fleiss formula, the arc-sine formula, and the formula of Casagrande, Pike, and Smith[3] for different assumed risk figures for P1:

P1	P2	Fleiss	Arcsine	Casagrande
4.0	4.4	65 433	65 965	75 831
3.0	4.4	4 691	4 872	5 690
4.1	4.4	117 677	118 376	135 884
2.5	4.4	2 357	2 504	2 950

These calculations show that if chorionic villus biopsy increases the spontaneous abortion rate by 0.4%, which would be equivalent to the risk for second-trimester amniocentesis, about 69000 pregnancies would be required in each group. The background rate of spontaneous abortion in the first trimester strongly influences the required numbers of patients—e.g. a drop to about 2600 patients in the two groups if the difference in abortion rates is about 2%. Even though the numbers required to achieve statistical significance are large, a study with matched controls allows a more meaningful statement about the added risk of spontaneous abortion after chorionic villus biopsy than the mere comparison with fetal loss rates in ultrasonically normal pregnancies now available. Only a well-designed, statistically sound, multicentre (preferably international) study can answer the very important questions about the safety of chorionic villus sampling.

W. HOLZGREVE Women's Clinic, Dept of Biomedical Statistics and Institute of Human Genetics, Westphalian Wilhelma University Munster, Germany

[3] Fleiss JL Statistical Methods for Rates an Proportions, New York Wiley, 1973.

Questions on above letter by Holzgreve et al :

a Why do the authors propose a 2-sample study? i.e.why not compare the proportion, p_2, of fetal losses observed following cvb in a single sample of n_2 pregnancies, against a "background rate" of P_1=3.7 (assume that this 3.7 is the figure they would have obtained by combining data from the literature, consulting experts, etc.)?

b What form would the data-analysis of such a "one-arm" study take? Use a numerical example with n_2=500 to illustrate.

c Calculate the required sample size for such a "one-arm" study, using the same α and β as they did (cf Colton p161).

d What form will the data-analysis of the "two-arm" study proposed by the authors take? Use a numerical example with n_1=n_2=500 to illustrate.

e Calculate the required sizes n_1 and n_2 for this study that the authors propose (cf Colton p168). Use P_1=3.0 (3rd row of table) and the same α and β. Note that the sample sizes may differ somewhat depending on the method of analysis, and on the formula used.

f Assume that a study of this size has been done and that the observed losses were p_1=3.8% and p_2=4.3%. What do you conclude? Use language that is understandable to those who will need to understand it.

g In the now-completed Canadian collaborative trial of cvb, the investigators plan to analyze the difference in all fetal losses and so are using P_1=6.6% and P_2=9.5% in their calculations. They used α=0.05 and β=0.20. What impact do these design differences have on sample size (full calculations are not required)?

-4- Analysis of Matched case-control study

In a case-control study, 317 patients suffering from endometrial carcinoma were individually matched with 317 other cancer patients in a hospital and the use of oestrogen in the six months prior to diagnosis was determined. The results were:

		Controls	
		Oestrogen used	Oestrogen not used
Cases	Oestrogen used	39	113
	Oestrogen not used	15	150

a What summary parameter can one estimate from these type of data?

b What is the point-estimate of this parameter?

c Derive a 95% CI for the parameter.

d perform a 2-sided test of significance to test the null hypothesis of no association between the oestrogen use and development of endometrial carcinoma.

-5- Analysis of un-matched case-control studies

A 1982 Swedish study (Arch. Env. Health, March/April 1982, p.81-) examined the association between exposure of female physiotherapists to non-ionizing radiations (shortwaves, microwaves,.) and the risk in subsequently delivered infants of a serious malformation or perinatal death. The exposures of two groups of working physiotherapists were compared: (a) the 33 mothers of the (33) infants who were born with serious malformations or who died perinatally; and (b) the (66) mothers of 66 randomly chosen "normal" infants. The resulting data, presented in a somewhat simplified form for this exercise, are:

Shortwave Use			Microwave Use *	Group	
	(a)	(b)		(a)	(b)
never/seldom	24	54	never	29	63
often/daily	9	9	sometimes	4	0

[* data missing on 3 mothers in group b]

a What summary parameter can one estimate from these types of data?

b What is the point-estimate of this parameter (analyze each exposure separately)?

c How WOULD you derive a CI for the parameter? (calculation not necessary).

d Perform a 2-sided test of significance to test the null hypothesis of no association between each of the two exposures and the subsequent delivery outcome.

-6- A SIMPLE WAY TO IMPROVE THE CHANCES FOR ACCEPTANCE OF YOUR SCIENTIFIC PAPER

To the Editor: During the past few years we have witnessed a revolution in the way manuscripts, abstract, and grant proposals are being typed. With improved typewriters and computer programs it is possible to produce manuscripts of typeset quality. It is generally assumed that data should be judged by its scientific quality and that this judgment should not be influenced by typing style.

I challenged this premise by analyzing the rate of acceptance of abstracts by a large national meeting. All abstracts submitted to the 1986 annual meeting of the American Pediatric Society and the Society of Pediatric Research (APS/SPR) appeared in Volume 20, No. 4 (Part 2) (April 1986) of *Pediatric Research.* Contrary to the practice of many other meetings, this volume also includes all the abstracts that were not accepted for presentation, and accepted papers are identified by symbols.

Abstracts were defined as "regularly typed" or "typeset printed." Each abstract was categorized as accepted if chosen for presentation or rejected.

A total of 1965 abstracts were evaluated. Excluded were 47 abstracts assigned for joint internal medicine-pediatric presentation, because the majority of them were submitted to the meeting of the American Federation for Clinical Research, and there was no indication of their rejection rate; only those that had been accepted appeared in the APS/SPR book of abstracts.

Of the 1918 evaluable abstracts, 1706 were regularly typed and 212 were "typeset." The acceptance rate was significantly higher for the "typeset" abstracts: 107 of 212 (51.4 percent) vs. 747 of 1706 (44 percent) (P<0.05).

Eighty-eight investigators submitted five or more abstracts to the meeting. Here, too, there was a higher rate of acceptance for the "typeset" abstracts (62 of 107:57.9 percent) as compared with the regularly typed abstracts (184 of 451:40.8 percent) (P = 0.002).

One may argue that investigators who can afford the new equipment for printing abstracts have more money and can afford better research, and therefore that their abstracts are accepted at higher rates. To explore this possibility. I analyzed data on the 15 investigators who submitted five or more abstracts each and who used both typing methods. In this subgroup, 19 or 55 regularly typed abstracts were accepted (34.5 percent), whereas 31 of 53 of the "typeset" abstracts were accepted (58.5 percent) (P = 0.015).

These results demonstrate that the new "typeset" appearance of data increases the chance of acceptance. It may mean that "typeset" printing may cause the data to look more impressive. Alternatively,it may mean that the new printing makes it easier for reviewers to read the data and to appreciate its meaning.

Most important, it means that this technological innovation reduces the chance of success of those not currently using it.

Questions

a. Display the data in the 5th paragraph in a 2 x 2 table.

b. What test (and what hypotheses) are appropriate to compare the "107 of 212 vs. 747/1706"? Notice that p<0.05. (Paragraph 5

c,d,e. see after rebuttal below

...ACCEPTANCE OF ABSTRACTS - A REBUTTAL

To the Editor: Dr. Koren claims that the use of a new "typeset" method for preparing an abstract may improve the chances for its acceptance at a national meeting, specifically, at the 1986 annual meeting of the American Pediatric Society and the Society for Pediatric Research (Nov 13 issue). This assertion, if correct, should raise alarm among investigators submitting their work for peer review and seeking a fair and objective critique. Although Dr. Koren lists several possibilities to explain why typeset printing may enhance the rate of acceptance of an abstract, including the possibility that printing may make the data appear more impressive or may make the reading of an abstract easier, his data can be interpreted differently.

Koren reports that 107 of 212 "typeset-printed" abstracts were accepted, as compared with 747 of 1706 "regularly typed" abstracts, the relative acceptance rates being 51.4 versus 44 percent (P<0.05). Because of the disparity in the sizes of the groups, we are uncertain what form of statistical analysis he employed. If one uses the technique of hypothesis testing of the differences between two proportions, the proportions 107 of 212 versus 747 of 1706 have a z value of 1849 with P<0.06. Thus, when an appropriate statistical method is used, a significant difference between the two proportions is not found at the 0.05 level.

These data can be examined in another way: 107 of a total of 854 accepted abstracts (12.5 percent) were "typeset," whereas 212 of 1918 abstracts submitted (11.1 percent) were "typeset." The difference between these proportions is obviously not significant. The difference in the sizes of the groups also makes it difficult to compare them. Furthermore, some abstracts were judged independently of this process in order to be placed in a poster symposium dealing with a specific topic (ie, "AIDS in Pediatric Patients"). Of the 30 abstracts chosen for these poster symposia, 15 were (we think) "typeset printed" and may appropriately be removed from the pool of accepted "typeset" abstracts.

Most important, a reviewer is judging the merit of a given abstract from a photocopy of the actual abstract, not its appearance in the April 1986 issue of *Pediatric Research.* "Typeset" abstracts that appear impressive in the abstract book do not necessarily stand out on the actual abstract form.

For these reasons, Koren's conclusion that a "technological innovation reduces the chance of success of those not currently using it" may not be entirely correct. Other reasons can be advanced to account for the apparent success of "typeset" abstracts.

Finally, in order to ensure that objective criteria are being used, all reviewers of abstracts for the 1987 meeting will receive a copy of Dr. Koren's letter so that they are aware of this potential problem.

R W. Chesney, M.D. Society for Pediatric Research University of California

Questions (continued)

c. The rebuttal claims that the difference between these two proportions is associated with a p-value of p=0.06 (2nd paragraph).

   Why do you think the "rebutting" authors arrive at a different p-value? [The typographical error (1819 for 1.849) is not the problem] (Paragraph 2, last two sentences)

d. In the 3rd paragraph of the reply, the authors look at the data regarding the same 1918 abstracts "in another way" i.e. in a type of case-control analysis. This is a legitimate way to look at the data; however, the "obviously nonsignificant" p-value associated with the comparison of 107/854 vs 212/1918 is not legitimate. Why? (Paragraph 3, fourth line)

e. The rebuttal mentions "the disparity in the sizes of the groups" in two places. The second time, in paragraph 3, it is stated that "the difference in the sizes of the two groups also makes it difficult to compare them". (Third paragraph, fifth line)Do you agree? Why / Why not?

-7- Test of a proposed mosquito repellent

An entomologist carried out the following experiment as a test of a proposed mosquito repellent. Thirty-five volunteers had one forearm treated with a small amount of repellent and the other with a control solution. The subjects did not know on which forearm the repellent had been used. At dusk the volunteers exposed themselves to mosquitoes and reported which forearm was bitten first. In 10/35, the arm with the repellent was bitten first.

a. Make a statistical report on the findings.

b. How would you analyze the results if:

   (i) some arms were not bitten at all?

   (ii) some people were not bitten at all?

8 EAR-CANAL HAIR AND THE EAR-LOBE CREASE AS PREDICTORS FOR CORONARY-ARTERY DISEASE

(NEJM Nov. 15, pp1318-1318, 1984]

To the Editor: The ear-lobe crease has been demonstrated to be significantly associated with coronary-artery disease in specific populations.[1] Patterns of hair growth have previously been suspected as possible risk factors for coronary-artery disease.[2,3] We investigated both the ear-lobe crease and ear-canal hair -- the presence of one or more terminal hairs growing on the tragus or antitragus or from the external acoustic meatus (Fig. 1) -- in 43 men and 20 women (36 to 76 years of age; mean, 56.3) who underwent coronary cineangiography. Coronary-artery disease was defined as a 50 per cent or greater luminal narrowing of one or more coronary arteries. Standard chi-square methods were used for the 63 subjects, and the McNemar test was used for 22 age-matched and sex-matched men (mean age, 51.2) on the variables of ear-lobe crease and ear-canal hair.

The ear-lobe crease was found to be significantly associated with coronary-artery disease (n=63, X^2=11.1, df=1, P < 0.001, Table 1), and a significant difference was seen between men with and without coronary-artery disease in the presence of ear-canal hair (n=22, X^2=4.0, df=1, P < 0.05, Table 2) when age was controlled for. The combined presence of ear-canal hair and the ear-lobe crease was found to be significantly associated with coronary-artery disease (n=43, X^2=4.77, df=1, P < 0.05, Table 3). Moreover, combining the ear-lobe crease and ear-canal hair yielded the greatest sensitivity (90 per cent) and the lowest false negative rate (10 per cent).

Table 1. Chi-Square Analysis of the Ear-Lobe Crease (ELC) in 63 Men and Women with and without Coronary-ArteryDisease

	Coronary-Artery Disease	
ELC	Present	Absent
Present	28	4
Absent	15	16

Table 2. McNemar's Test of Ear-Canal Hair (ECH) in 11 Pairs of Age-Matched Men with and without Coronary-Artery Disease (CAD).

Distribution within Pair		No. of Pairs
CAD + and ECH +	CAD – and ECH +	6
CAD + and ECH +	CAD – and ECH –	4
CAD + and ECH –	CAD – and ECH +	0
CAD + and ECH –	CAD – and ECH –	1

Table 3. Chi-Square Analysis of the Ear-Lobe Crease (ELC) and Ear-Canal Hair (ECH) in 43 Men with and without Coronary-Artery Disease.

	Coronary-Artery Disease	
ELC & ECH	Present	Absent
Present	18	2
Absent	14	9

The frequency of hairy pinnae in men varies according the genetically defined populations, and the penetrance of this trait is variable.[4] Various amounts of hair may grow anywhere on the external ear, and specific loci of hair growth are seen in specific populations.[4] Hairy pinnae are unusual in women,[5] and none were found in this study. Ear-canal hair was found to be present in 74.4 per cent of men in this study.

Androgens may facilitate the development of atherosclerosis and coronary-artery disease may be due to the long-term exposure to enough androgen to cause both ear-canal hair growth and coronary-artery disease. The degree of androgenicity in a patient over a period of years may explain the eventual virilization of the ear and the associated accelerated atherosclerosis in these patients. Another androgen-sensitive trait, male pattern baldness, has also been recognized as a predictor of coronary thrombosis in men, possibly on the same basis.

Richard F. Wagner, Jr., M.D., Howard B. Reinfeld, M.D., Karen Dineen Wagner, M.D., PhD., Anthony T. Gambino, M.D., Thomas A. Falco, M.D., Jerry A. Sokol, M.D., Stanley Katz, M.D., and Steven Zeldis, M.D. Nassau Hospital

Questions

(a) Why did the authors consider it important to use an age-matched comparison when studying ear-canal hair but an unmatched comparison for ear-lobe crease?

(b) Verify the X^2 of 11.1 in Table 1.

(c) Reconstruct the X^2 of 4.0 in Table. 2. What would the p-value be if the authors had used the binomial table rather than the X^2 table for Table 2? Can you reconcile the difference in the 2 p-values?

(d) Do you agree with the authors' choice of analysis and interpretation of the data in Table 3?

(e) Comment on their statement regarding the sensitivity and false negative rate of the ear-lobe crease/ear-canal hair combination.

(If you wish, write your answers to (b)-(e) in the form of the Letter to the Editor)

-9- Windsurfing data

Carry out overall and trend test on windsurfing data (given above in §8.3)

-10- Right-Handedness: A consequence of Infant Supine Head-Orientation Preference?

Most newborn infants orient their heads towards their right sides while supine. This right bias has been thought to contribute to the development of right bias in handedness by producing lateral symmetries in visual experience of the hands and differences between the hands in neuromotor activity. In a study to investigate this theory, some 150 neonates were assessed in the 16 to 48 hours after birth, resulting in the following distribution of neonatal head-orientation preference

Definitely Right	Right Tendency	Mixed	Left Tendency	Definitely Left	Total
73	24	31	13	9	150

leading the author to conclude that the distribution was "significantly biased to the right".

Twenty neonates with consistent head-orientation preferences were selected from the original 150 (10 from each extreme) and tested at 22 weeks for hand use preference, giving the following results: (R = Right; L = Left).

10 infants who consistently oriented head to **right**

Neonatal Head-Orientation	Hand-Use Preference at 22 Wks	
	Initial Reach	Frequency Score
R	R	1.0
R	R	0.4
R	R	2.0
R	R	1.2
R	*	0.2
R	L	-2.5
R	R	1.5
R	R	1.3
R	R	1.9
R	R	1.9

10 infants who consistently oriented head to **right**

Neonatal Head-Orientation	Hand-Use Preference at 22 Wks	
	Initial Reach	Frequency Score
L	L	-2.3
L	L	-2.3
L	*	0.0
L	L	-1.4
L	R	1.3
L	L	-1.9
L	L	-2.3
L	L	-1.0
L	L	-1.0
L	R	1.8

* Each hand was used for initial reaching in half the testing conditions

Questions

a Do you agree that the distribution of head-orientation preferences is "significantly biased to the right"? How would you put it to a statistical test?

b Does the direction of neonatal head orientation significantly predict which hand is used initially in a 3 minute test? To think about this, it might help to imagine trying to predict hand preference from whether the baby was born on an even or odd day of the month.

c What about its ability to predict reaching frequency preference? (a positive frequency score means the infant reached more often with the right hand during the full 3-minute test; a negative score meant he/she reached more often with the left.)

d The author claims that "infants with consistent preferences to turn their heads to the right show a significant right-hand bias (as judged by positive frequency scores) at 22 weeks (bionomial sign test, p = 0.0215, two-sided). Explain how this p-value was obtained; judge whether infants with a left head orientation preference are similarly biased towards left-handedness?

-11- Triangle Taste test

In its 1974 manual "Laboratory Methods for Sensory Evaluation of Food", Agriculture Canada described tests (the triangle test, the simple paired comparisons test,...) to determine a difference between samples.

"In the triangle test, the panelist receives 3 coded samples and is told that 2 of the samples are the same and 1 is different and is asked to identify the add sample. This method is very useful in quality control work to ensure that samples from different production lots are the same. It is also used to determine if ingredient substitution or some other change in manufacturing results in a detectable difference in the product. The triangle test is often used for selecting panelists.

Analysis of the results of triangle tests is based on the probability that - IF THERE IS NO DETECTABLE DIFFERENCE - the odd sample will be selected by chance one-third of the time. Tables for rapid analysis of triangle test data are given below. As the number of judgements increases, the percentage of correct responses required for significance decreases. For this reason, when only a small number of panelists are available, they should perform the triangle test more than once in order to obtain more judgements.

The results of a test indicate whether or not there is a detectable difference between the samples. Higher levels of significance do not indicate that the difference is greater but that there is less probability of saying there is a difference when in fact there is none"

Chart: Triangle test difference analysis
[Table starts at n=7 and ends at n=2000; selected entries shown here]

Number of Tasters	Number of correct answers necessary to establish level of significance		
	5%	1%	0.1%
7	5	6	7
10	7	8	9
12	8	9	10
30	16	17	19
60	28	30	33
100	43	46	49
1000	363	372	383

Questions

a Show how one arrives at the numbers 7, 8 and 9 of correct answers necessary to establish the stated levels of significance for the case of n=10 tasters. Hint: you can work them out from the BINOMDIST function in Excel or [since we are only interested in the principles involved, and not in getting answers correct to several decimal places] you should be able to interpolate them from probability distributions tabulated in the text [the setup here is similar to the therapeutic touch study, but with p=1/3 rather than p=1/2].

b Calculate the exact 90%, 98% and 99.8% 2-sided CI's for the proportions 7/10, 8/10 and 9/10 respectively, and from these limits verify that indeed 7/10, 8/10 and 9/10 are significantly greater than 0.33, at the stated levels of significance .(I am presuming that their H_a is 1-sided, ie. 0.33 vs. > 0.33)

You can obtain these CI's from the spreadsheet "CI for a proportion", under Resources for Ch 8.

c Show how one arrives at the numbers 43, 46 and 49 of correct answers necessary to establish the levels of significance for the case of 100 tasters. Hint: you should be able to use a large-sample approximation.

d How well would this large-sample approximation method have done for the case of n=10?.

e If you set the alpha at 0.05 (1-sided), what number of tasters is required to have 80% power to 'detect' a 'shift' from H_0: p=1/3 to (i) H_a: p=1/2 (ii) H_a: p=2/3? Use the sample size formula in section 8.1 of the notes.

Notes: See worked example 2 in notes on Chapter 8.1. This is an good example where a one-sided alternative is more easily justified, so with α = 0.05 1-sided, Z_α = 1.645. Note that power of 80% means that β = Prob(failing to reject H_0) = 1 – β, so Z_β = -0.84. The Z_β is always one-sided, since one cannot be on both sides of H_0 simultaneously!

f "The triangle test is often used for selecting panelists." -- end of ¶2. Presumably, if one had to choose one of two available panelists, one would ask each to make several judgements. How many judgements would you ask each to make? State any assumptions you make .

g "When only a small number of panelists are available, they should perform the triangle test more than once in order to obtain more judgements" -- end of ¶3. What scientific objection might one have to this advice?

h Do you agree with the statement "Higher levels of significance do not indicate that the difference is greater but that there is less probability of saying there is a difference when in fact there is none"--end of ¶4. Why ?

i Explain to somebody who knows little statistics why you think a study with n = 6 tasters would not tell very much. Be statistical, but avoid jargon like 'power' and 'significance' and 'hypothesis'.

j With a small n umber of testers, it is possible that, even if a sizeable proportion of the population can correctly taste the Δ, the test of significance will be 'negative'. Suppose that 50% can truly tell the Δ and that 1/3 of the remaining 50% get the test correct by guessing, giving an overall 67% who get the test correct. In this situation, what is the probability that a trial with n=12 will yield a 'positive' (i.e. statistically significant) answer? What if the trial uses n=30? n=60?

-12- More U.S. PhD's At McGill than Canadian

MCGILL DAILY, 1993.09.08

McGill professors with a doctorate from Canada are a rare breed when compared to their colleagues who were educated in the United States. According to the 1993-1994 Calendar, in the Faculties of Arts and Science, 42% of professors have American PhDs whereas only 36% have Canadian. This trend has worried some who feel that Canadian PhD graduates are being discriminated against by Canadian universities, and that an education in the United States is unfairly valued over one obtained in Canada.

Letter To MCGILL DAILY September 9, 1993

Considering the numerous issues of real importance that exist, why do you have to invent more? I am referring to your September 8 front page article "More U.S. PhDs at McGill than Canadian," the first sentence of which reads "McGill professors with a doctorate from Canada are a rare breen when compared to their colleagues who were educated in the United States." The second sentence contradicts this; it points out that 36% of Arts and Science professors have Canadian PhDs, vs 42% with U.S. PhDs. This is a deviation of only 6%: roughly the margin or error of Gallop polls. Those who claim that Gallop polls have margin of error of only 4% have forgotten the necessary multiplication by the square root of two. A roughly one to one ration hardly makes Canadian PhDs a "rare breed." In fact, according to your statistics, over one third of our professors... [underlining mine... jh]

Comment and Questions:

The letter writer asks the Daily «why in your first sentence do you use the phrase "rare compared to their colleagues" when the percentages are 36 and 42? »[1]

We could ask the letter writer «why do you use the phrase "deviation of only 6%" when a simpler "difference of only 6%" would do equally well» and «why complicate things by mentioning Gallop polls and margins of error and the square root of two?»

a In *one sentence*, explain why one doesn't need inferential statistics here.

b Also, explain to this writer that if (s)he is going to bring statistical inference about proportions into this, (s)he should get his margin of error correct.

-13- Women faster drivers: survey MONTREAL GAZETTE, 20/2/95

LONDON - Woman drivers in Britain are more likely than men to exceed the speed limit, according to a survey by Autoglass, an international supplier of replacement glass based in London. Paul Eyton-Jones, marketing manager of Autoglass, said the survey examined the driving habits of 400 people as a means of improving road safety. The results showed that 21 per cent of women exceed the speed limit of 70 mph compared with 19 per cent of men. Only 14 per cent of women would drive at the safer speed of 60 mph compared with 38 per cent of men. "We've always thought it's the men drivers, the ego, pushing up the fast lane," Eyton-Jones said. "What we found is that women are exceeding the limit in equal measure."

Questions:

a There is not enough information here to judge the study design. What main features would you be looking for when you read the Methods Section of the full report?

b Assuming that you found the design to be good, carry out a formal test of the 21% vs. 19% exceeding 70mph (113Km/h)

c You do not have the numbers of men and women studied, so you assume it was 200 of each. If your assumtion is not correct (say the real numbers were 300 women and 100 men), how will the p-value you calculate compare with the one using the correct numbers?

while we are on the topic...

-14- WOMEN ARE SAFER PILOTS: STUDY

LONDON- Initial results of a study by Britain's Civil Aviation Authority shows that women behind the controls of a plane might be safer than men. The study shows that male pilots in general aviation are more likely to have accidents than female pilots. Only 6 per cent of Britain's general aviation pilots are women. According to the aviation magazine Flight International, there have been 138 fatal accidents in general aviation in the last 10 years, and only two involved women - less than 1.5 per cent of the total.

WomanNews, page F1

[1] The writer should also complain about the use of "compared to" ; the correct usage is "compared with".

The Montreal Gazette, August 21st, 1995

Questions:

a What is the comparative parameter at issue here?

b Comment on the epidemiologic soundness of the comparison reported.

c Assuming that the comparison reported is a sound one, or that it can be made so using additional information, translate the data into point and interval estimates of the comparative parameter. Also, carry out a test of the null value of the comparative parameter.

-15- Perioperative Normothermia

Refer to the report of this study (scanned version of text as images [.gif files] under Resources for Chapter 5; full version, using optical character recognition, and reformatting in a word processor, as a pdf file in Resources for Chapter 7)

a Using the same 'inputs' as the authors did (2nd paragraph of Methods), calculate the sample size requirements.

Some formulae do not use different null and non-null variances, instead, for simplicity, they use the same null and non-null variance --calculated at the average of the null and non-null p's; and some authors use a formula based not on the difference of the proportions, but of the arcsine transformations of these proportions. Thus, you should not be surprised if you don't get exactly the same numbers.

See also my footnote concerning the choice of 'delta'. The difference that would be important (the clinically important difference) is a matter of judgment; it should not be left to be 'dictated' empirically by Nature (the authors used as their 'delta' the empirical difference 9/38 - 4/42 = 14.2% found in their pilot study!). Imagine what the authors' 'delta' could gave been if they had done a pilot study of say 2 patients vs. 3 patients, or just 1 vs. 2! And , even with increasing sample sizes, Nature is just going to show you more precise estimates of what the difference is, not of "the difference that would make a difference". After all, Nature doesn't know how much these normothermia blankets cost, or how acceptable and practical they would be!

Indeed, it is ironic that the observed difference in the study proper is only 19% - 6% = 13%; it is "statistically significant" but less than the 'clinically important delta' used by the authors in their sample size formula.

b State the null and alternative hypotheses, and re-calculate the P-value in the first row of Table 2.

c Calculate a 95%CI for the difference in infection rates.

d You can convert the point estimate of the difference into the "number required to treat". The formula for this is

$$1/(\text{Infection Rate}_{\text{if do not treat}} - \text{Infection Rate}_{\text{if treat}})$$

The logic is that if 19/100 would develop an infection without the intervention, and 6/100 despite it, then intervening on 100 would prevent 19 - 6 = 13 infections, i.e.. one would need to intervene on approximately 8 (i.e. 100/13) to prevent 1 infection.

Convert the upper and lower 95% limits for the difference (from part c) into the corresponding limits on the number required to treat.