1. Questions re van Belle et al.

i. From Problems 6.1 to 6.30, find 1 that involves (i) a test of a single proportion (ii) a CI for a single proportion (iii) a test of the equality of two proportions, a CI for (iv) a RD (v) a Risk Ratio and (vi) an Odds Ratio.

ii. Problem 6.1(b) raises a subtle and important point, but does not say why the requested probability calculation helps in evaluating the complaint. Explain what is the appropriate probability to calculate in order to judge if the clinic's complaint is valid.

-2- Sample size to assess 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" or "parameter") symbols "π_1" and "π_1". 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, i.e., the 'statistic.'

> 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 (P_2) we have calculated the required sample size for the two populations, based on a probability of a type I error (α) of 1% and of a type II error (b) of 10%. The most recent international survey^{ref} revealed a spontaneous abortion rate of about 4.4% after chorionic villus sampling, and this was the figure we used for the rate in P_2 when calculating sample sizes by the Fleiss formula, the arc-sine formula, and the formula of Casagrande, Pike, and Smith[1] for different assumed risk figures for P_1:
>
> | P_1 | P_2 | Fleiss | Arcsine | Casagrande |
> |---|---|---|---|---|
> | 4.0 | 4.4 | 654,33 | 65,965 | 75,831 |
> | 3.0 | 4.4 | 4,691 | 4872 | 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 69,000 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 2,600 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 Biomed. Statistics & Inst of Human Genetics, Westphalian Wilhelma Uni., Munster, Germany.

Questions on above letter:

i. 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.

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

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

iv. 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.

[1] Fleiss JL Statistical Methods for Rates an Proportions, 1973.

v. Calculate the required sizes n_1 and n_2 for this study that the authors propose (cf Notes, or vanBelle, or 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.

vi. 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.

vii. 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 $\alpha = 0.05$ and $\beta = 0.20$. What impact do these design differences have on sample size? Full calculations are not required.

-3- 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. Two *series* of working physiotherapists were compared: ($Y = 1$) the 33 mothers of the (33) infants who were born with serious malformations or who died perinatally; and ($Y = 0$) the (66) mothers of 66 randomly chosen "normal" infants. The resulting data, presented in a somewhat simplified form for this exercise, are:

	Y			Y	
Shortwave Use	1	0	*Microwave Use**	1	0
never/seldom	24	54	never	29	63
often/daily	9	9	sometimes	4	0

* data missing on 3 mothers for whom $Y = 0$.

i. What comparative parameter can one estimate from these data? Think of the $Y = 1$ data as coming from the *numerator series*; think of the $Y = 0$ data as coming from the *denominator series* that supplies *estimates* of the fractions of the source population that are in the higher- and lower-use categories.

ii. For each the two exposures, what is the point-estimate of this parameter?

iii. Derive a 95% CI for the parameter, by "Woolf's" method for shortwave, the exact conditional method (Fisher) for microwave (see spreadsheet).

iv. 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.

4. 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, abstracts, 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

i Display the data in the 5th paragraph in a 2×2 table.

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

iii-v 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)

iii 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)

iv 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" pvalue associated with the comparison of 107/854 vs 212/1918 is not legitimate. Why? (Paragraph 3, fourth line)

v 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?

-5- 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.

i. Make a statistical report on the findings.

ii. How would you analyze the results if:
- some arms were not bitten at all?
- some people were not bitten at all?

-6- 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)

i. 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 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.

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

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

iv. You can convert the point estimate of the difference into the "number required to treat." The formula for this is 1/(Infection Rate if Do Not Treat - Infection Rate 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 iii) into the corresponding limits on the number required to treat.

-7- 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.

[Montreal Gazette, WomanNews, page F1]

i. What is the comparative parameter at issue here?

ii. Comment on the epidemiologic soundness of the comparison reported.

iii. 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.

-8- Equivalent Forms of the X^2 statistic from a 2×2 table
Consider a 2×2 table with frequencies $y_1 = a,\ b,\ y_2 = c,\ d$, row totals $n_1 = r_1,\ y_2 = r_2$, column totals $c_1,\ c_2$, overall total n, observed proportions $p_1 = y_1/n_1$ and $p_2 = y_2/n_2$, overall proportion $p = (y_1+y_2)/n$, and $Var[a|H_0]$ based on the '2-independent-binomials' model. Show that

$$\begin{aligned} X^2 &= \sum \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency} \\ &= n \times \frac{(a \times d\ -\ b \times c)^2}{r_1 \times r_2 \times c_1 \times c_2} \\ &= \frac{\{p_1 - p_2\}^2}{p(1-p) \times (1/n_1 + 1/n_2)} \\ &= \frac{\{a - \widehat{E[a \mid H_0]}\}^2}{Var[a \mid H_0]}. \end{aligned}$$

-9- Bone mineral density and body composition in boys with distal forearm fractures

J Pediatr 2001 Oct;139(4):509-15.

Goulding et al (New Zealand)

Abstract

Objective: To determine whether boys with distal forearm fractures differ from fracture-free control subjects in bone mineral density (BMD) or body composition. **Study design**: A case-control study of 100 patients with fractures (aged 3 to 19 years) and l00 age-matched fracture-free control subjects was conducted. Weight, height, and body mass index were measured anthropometrically. BMD values and body composition were determined by dual-energy x-ray absorptiometry. **Results**: More patients than control subjects (36 vs l4) were overweight (body mass index > 85th percentile for age, P < .001). Patients had lower areal (aBMD) and volumetric (BMAD) bone mineral density values and lower bone mineral content but more fat and less lean tissue than fracture-free control subjects. The ratios (95% CIs) for all case patients/control subjects in age and weight-adjusted data were ultradistal radius aBMD 0.94 (0.91-0.97); 33% radius aBMD 0.96 (0.93-0.98) and BMAD 0.95 (0.91-0.99); spinal L2-4 BMD 0.92 (0.89-0.95) and BMAD 0.92 (0.89-0.94); femoral neck aBMD 0.95 (0.92-0.98) and BMAD 0.95 (0.91-0.98); total body aBMD 0.97 (0.96-0.99), fat mass 1.14 (1.04-1.24), lean mass 0.96 (0.93-0.99), and total body bone mineral content 0.94 (0.91-0.97). **Conclusions**: Our results support the view that low BMC, aBMD, and BMAD values and high adiposity are associated with increased risk of distal forearm fracture in boys. This is a concern, given the increasing levels of obesity in children today. (J Pediatr 2001;139:509-15)

		Fracture?	
		Yes	No
Overweight?	Yes:	36	14
	No:	64	86
	Total	100	100

i. Rewrite the sentence "A case-control study of 100 patients with fractures (aged 3 to 19 years) and l00 age-matched fracture-free control subjects was conducted" using terminology that better reflects the purpose of the 100 fracture-free subjects.

ii. All of the fractures occurred over a 1-year period, ten of them in persons aged 11. Suppose one could choose a random sample, of size ten, from **all** 600 11-year old boys living in the city of Dunedin, what is the probability that this denominator series would have an overlap of 0, 1, 2, .. with the case series of ten? [2]

What if age-matching were to the nearest month of age, and that there were two cases in boys aged 11 years and 3 months, so we took a sample of two from all of the 600?

iii. Estimate the ratio of the fracture rate in the overweight to the fracture rate in the not-overweight, and use Woolf's method to calculate a 95% CI for it (ignore the age-matching).

iv. We can repeat the point- and interval estimation using logistic regression: e.g., in `R`,

```
y=c(rep(1,100),rep(0,100)); over=c(rep(1,36),rep(0,64),rep(1,14),rep(0,86))
summary(glm(y~over,family=binomial))
```

yielding...

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2955     0.1651  -1.790 0.073490 .
over          1.2399     0.3556   3.487 0.000489
```

Verify that 1.2399 represents the $\log or$ and 0.3556 its SE.

[2] In fact, the "age-matched denominator series" was assembled as follows: All patients with fractures were asked to supply the names of 3 friends of their own age: the first friend who had never fractured a bone at any time of his life and who agreed to take part as a fracture-free control subject was then enrolled."

-10- Theoretical basis for "odds ratio" as estimator of Rate Ratio, together with statistical model for the estimator

The old-fashioned and very loose justification for using the empirical odds ratio, *or*, as an estimator of the theoretical rate ratio goes back to Cornfield in the 1950s. Unfortunately it still is the one given in many 'modern' texts, despite the much more general justification provided by Miettinen in 1976.

The old justification rested on algebraic arguments using *persons*, not *population time*. The outcome *proportions* involved refer to *cumulative* incidence.

The truly modern way is to think of the cases as arising in population-time, and to think of the population time involved as an infinite number of person-moments - think of a person-moment as a person at a particular moment. Say that a proportion π_E of these are "exposed" person moments, and the remaining proportion π_0 are "non-exposed" person-moments. Suppose further that the (theoretical) event rates in the exposed and unexposed amounts of population-time are

$$\lambda_E = \frac{E[no.events]}{PT_E} \; ; \; \lambda_0 = \frac{E[no.events]}{PT_0},$$

with (theoretical) Rate Ratio $\theta = \lambda_E/\lambda_0$.

Denominator Series

Suppose we take a finite random sample, of size d, of the infinite number of person moments in the base that generated the cases, and classify them into d_E "exposed" person moments and $d_0 = d - d_E$ "non-exposed" person-moments. We will refer to this sample of d as the *denominator* series. What is the statistical model for $d_E \mid d$? Clearly, it is

$$d_E \sim Binomial(d, \pi_E).$$

Numerator (Case) Series

Denote by c the observed number of events; we classify them into c_E events in "exposed" population-time and $c_0 = c - c_E$ in the "non-exposed" population-time. We will refer to this sample of c as the *case* series.

What is the statistical model for $c_E \mid c$? We can think of c_E as the realization of a Poisson r.v. with mean (expectation) $\mu_E = (PT_E \times \pi_E) \times \lambda_E$. Likewise, think of for c_0 as the realization of a Poisson r.v. with mean (expectation) $\mu_0 = (PT_0 \times \pi_0) \times \lambda_0$.

Now, it is a statistical theorem (Casella and Berger, p194, exercise 4.15) that

$$c_E \mid c \sim Binomial(c, \mu_E/[\mu_E + \mu_0]).$$

Thus we can identify the distribution of the 4 random variables involved in the OR estimator

$$\hat{OR} = or = c_E/d_E \; \div \; c_0/d_0 \; = \; c_E/c_0 \; \div \; d_E/d_0 = (c_E \times d_0) \; \div \; (c_0 \times d_E).$$

The $c_E : c_0$ split is governed by one binomial, involving θ and other parameters, while the $d_E : d_0$ split is governed by a separate binomial, involving the same other parameters, but not involving θ.

If one replaces μ_E and μ_0 by their constituents, one can show that the odds that an unexposed person-moment in the series of $c + d$ represents a "case" is $c : d$, whereas the corresponding odds for an exposed person moment is $(\theta \times c) \ : \ d$.

In other words, in the dataset of $c + d$,

$$logit[Prob[case|0] = \log(c/d) = \beta_0 \ ; \ logit[Prob[case|E] = \log(c/d) + \log\theta = \beta_0 + \beta_E E,$$

where E is an indicator variable.

So, one can estimate $\log\theta = \log OR$ by a logistic regression of the $c + d$ Y's i.e. $Y = 1$ if in case series; $= 0$ if in denominator series, on the corresponding set of $c + d$ indicators of exposure (1 if exposed, 0 if not).