8.47 A major court case on the health effects of drinking contaminated water took place in the town of Woburn, Massachusetts. A town well in Woburn was contaminated by industrial chemicals. During the period that residents drank water from this well, there were 16 birth defects among 414 births. In years when the contaminated well was shut off and water was supplied from other wells, there were 3 birth defects among 228 births. The plaintiffs suing the firm responsible for the contamination claimed that these data show that the rate of birth defects was higher when the contaminated well was in use.[14] How statistically significant is the evidence? What assumptions does your analysis require? Do these assumptions seem reasonable in this case?

8.75 An experiment designed to assess the effects of aspirin on cardiovas
disease studied 5139 male British medical doctors. The doctors we
randomly assigned to two groups. One group of 3429 doctors took
aspirin daily, and the other group did not take aspirin. After 6 years
were 148 deaths from heart attack or stroke in the first group and 79
the second group. A similar experiment used male American medi
doctors as subjects. These doctors were also randomly assigned to o
two groups. The 11,037 doctors in the first group took one aspirin
other day, and the 11,034 doctors in the second group took no aspi
After nearly 5 years, there were 104 deaths from heart attacks in the
group and 189 in the second.[27] Analyze the data from these two studi
summarize the results. How do the conclusions of the two studies c
and why?

8.81 *Castaneda v. Partida* is an important court case in which statistical methods were used as part of a legal argument.[28] When reviewing this case, the Supreme Court used the phrase "two or three standard deviations" as a criterion for statistical significance. This Supreme Court review has served as the basis for many subsequent applications of statistical methods in legal settings. (The two or three standard deviations referred to by the Court are values of the z statistic and correspond to P-values of approximately 0.05 and 0.0026.) In *Castaneda* the plaintiffs alleged that the method for selecting juries in a county in Texas was biased against Mexican Americans. For the period of time at issue, there were 181,535 persons eligible for jury duty, of whom 143,611 were Mexican Americans. Of the 870 people selected for jury duty, 339 were Mexican Americans.

(a) What proportion of eligible voters were Mexican Americans? Let this value be p_0.

(b) Let p be the probability that a randomly selected juror is a Mexican American. The null hypothesis to be tested is H_0: $p = p_0$. Find the value of $\hat{p}$ for this problem, compute the z statistic, and find the P-value. What do you conclude? (A finding of statistical significance in this circumstance does not constitute a proof of discrimination. It can

be used, however, to establish a prima facie case. The burden of pr then shifts to the defense.)

(c) We can reformulate this exercise as a two-sample problem. Here w wish to compare the proportion of Mexican Americans among tho selected as jurors with the proportion of Mexican Americans amor those not selected as jurors. Let p_1 be the probability that a rando selected juror is a Mexican American, and let p_2 be the probabilit that a randomly selected nonjuror is a Mexican American. Find th z statistic and its P-value. How do your answers compare with you results in (b)?

-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?

-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?

_____

-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.