Department of Epidemiology and Biostatistics, McGill University
EPI 513-607 (Inferential Statistics) Final Examination
June 23, 1992

marks out of 205 (items mostly out of 5, a few out of 10 are marked with a # sign)

Article: Reduction of transmission of shigellosis by control of houseflies by Cohen et al Lancet April 27, 1991

Materials and methods -- study design (4th paragraph).

a* Sketch one other possible design.

	1988 I	1988 II	1989 I	1989 II
Base X	I	I	C	C
Base Y	C	C	I	I

	1988 I	1988 II	1989 I	1989 II
Base X	I	I	I	I
Base Y	C	C	C	C

b* Give one major advantage it would have over that the authors used.
c* Give one major disadvantage it would have.

Materials and methods -- bacterial culture of flies.

a If *shigella* sp were "isolate–able" in 1% of individual flies, what proportion of pooled samples of 10 flies would be expected to have shigella isolated from them? State any assumptions made.

If the 10 are a reasonable random sample of flies, and if the pool is positive when ≥ 1 of the 10 is positive, then the probability of a positive pool is given by summing the Binomial probabilities BinProb(X=x|n=10, π=0.01) for x=1,2,...10.

It is easier to do the calculation as 1 – BinProb(X= 0 | n=10, π=0.01) i.e. as 1 – $(0.99)^{10}$ or 1 – 0.904 i.e. as 0.096 {the 0.904 could also be obtained directly from the binomial table on page 2.19 of the notes}.

Materials and methods -- Clinical and laboratory surveillance.

a Why do you think the authors decided that *"the few individuals with more than one episode of diarrhoea were counted only once in the analyses"* ?

The episodes may not be randomly distributed over individuals, but may be overly clustered in certain individuals. One way is to count 1st episodes; another is to calculate the number of episodes per person and analyze it as a "continuous" variable (it will still be mainly 0's and 1's)

b *"A significant rise in antibody was defined as …"* If you were one of the authors, and a referee asked you to state what p-value you used for significance and whether you used a paired t-test or a non-parametric analog, how would you reply?

This is a clinical judgement and uses "significant" in the sense of substantial or important. I suppose that if one wanted to, one could define it statistically as "a change that is bigger than the measurement variation".

Materials and methods -- Statistical methods.

a What statistical distribution would you think of for the daily variation in fly counts [within 1 training session in 1 base]. Why? Why might it not fit that well?

For counts, the first one that comes to mind is the Poisson distribution. It is a distribution for counts (e.g. lightning strikes, mortality, etc) and the counts are from samples from a defined denominator (2.5 m of wire) . The requirement is that the population remain stable over time. Here one might expect the population of flies to vary over the months covered by one session i.e. there may be peak weeks. This would introduce additional variation beyond just the sampling variation. We can check this later.

b Why did the authors choose the Wilcoxon rank sum test for comparing fly counts between intervention and non-intervention base at each study session?

I expect that they examined the distribution (their n's were big enough to get a reasonable feel for the distribution) and saw that it was fairly non-Gaussian (possibly skewed, multimodal, etc.). They were probably afraid that the assumptions for a parametric test were not met.

c What parametric alternative was available? Is it seriously contraindicated here?

The t-test for independent samples. Even though the raw counts are probably quite non–Gaussian, the n's are large and the Central Limit Theorem should guarantee that the t or the z tables will be reasonably accurate.

d Why " χ^2 OR Fisher's exact test " ?

In some tables (e.g. comparison of diarrhoeal rates), the expected counts in the cells are large enough to comfortably use the χ^2 distribution , whereas in others (Shigellosis) they are not.

e Explain more fully the reasons for the "conservative statistical analysis". Imagine trying to convince a skeptic, who does not like to see the degrees of freedom reduced from the hundreds to the single digits, and the p-values rendered less impressive.

The randomization was by base. The session in one base should be viewed collectively since there are many influences (e.g. the cafeteria, other sources of illness) that act on the entire base; these may have nothing to do with flies but still act on all individuals. This 'cluster randomization' is much more susceptible to major variations than a randomization of individuals to personalized interventions to prevent individual (non-contagious) outcomes. One way to argue is to imagine that the study had been done in just one base in one year. Even if the numbers of soldiers were in the thousands, a statistically significant difference between sessions would simply mean that the difference was numerically non-zero, but there might be a number of competing explanations; 2 sessions in 1 base in 1 year would only contribute 1 degree of freedom, and so to attribute the observed difference to the intervention would be to deny all other explanations.

f Why, with their conservative approach, are the authors so much more worried about normality before carrying out Student's t tests? They transformed[#] the proportions; why did they not transform the raw mean fly counts?

There are now only 4 paired observations, so we would have to use a t_3 distribution. The t distribution assumes that the paired differences come from a Gaussian distribution of paired differences, something that is difficult to check with only 4 observations.

The transformation doesn't necessarily induce Gaussian–ness in the paired differences, although under H_0, the differences should at least have a symmetric distribution.

g Would you have recommended nonparametric tests instead? Why? Why not?

Yes, I would. See previous answer.

h The authors define "statistically significant" as "$p \leq 0.05$". Define "statistically significant" in words.

If all that was at work was random (sampling) variation, then the probability of obtaining an average difference as big as or bigger than that observed, is less than 5%. This is generally taken as a low enough probability that it casts doubt on the assertion that the only factor at work is random variation.

for those interested

"The mean rates (i.e. the proportions) for each base during each study session were then arcsine-transformed to better approximate normality".

Notes (almost verbatim from A&B): The motivation for this transform is variance stabilization {constancy of variance [for a fixed n] over the various values of x is a requirement for some inferences from regression}. It is called the angular, inverse sine or arcsine transformation and is calculated from the observed proportion p as $\sin^{-1}[\sqrt{p}]$ i.e. the angle whose sin is $\sqrt{p}$. Armitage gives the following short table for 100p (%) for %'s from 0 to 50 {the angles for %'s > 50 are obtained by subtracting the angle corresponding to 100–% from 90°}

100p % :	0	5	10	15	20	25	30	35	40	45	50
angle ° :	0	13	18	23	27	30	33	36	39	42	45

Equal changes in the % correspond to greater changes in the angle towards the two ends of the scale than near the middle; the arcsin transform of a binomial proportion has a variance of approximately 821/n , in contrast to the variance of p itself of $\pi\{1-\pi\}/n$, which varies from 0/n to 1/4n to 0/n as π varies from 0 to 0.5 to 1. The constant variance of 821/n for the arcsine transformed proportions may be used as a baseline in analyses to check whether residual variation is greater than binomial.

Results -- fly density.

a Translate the "$p<0.0001$" accompanying the r's of 0.79, 0.84 and 0.86 into natural language. Comment on the null hypothesis being tested.

If the 2 measures were uncorrelated, and if all that was at work was random (sampling) variation, then the probability of obtaining a correlation as far (or farther) away from zero as that observed, is less than 0.01%.

The (null) hypothesis of "zero correlation between measures" is not really of interest. We expect the correlation to be positive; the only question is how positive. If one thought that a correlation ρ were the appropriate measure here, then a CI for ρ would make more sense

b How would you have presented the data to show that *"the two techniques gave similar results "* ?

Much more descriptively: (i) graphically as a plot of the discrepancy in the paired counts versus the count by the gold standard method (or the average of the 2 if one doesn't consider the more complex method a more definitive method). Here, even the gold standard method produces error-containing estimates, since it involves sampling in space and time. (ii) by reporting the range of discrepancies.

Since very few students gave answers like these, I concluded that either (i) very few had done the assigned exercise on peak flow measurements (using the Bland and Altman paper in Lancet that I had on reserve) or (ii) most thought [rightly so] that the counts from the 2 methods would not necessarily be commensurate, since one used a grid and the other a line (i.e.. the volume counted was not the same). If this is the case, it would have made sense to use the ratio of the two counts ecah day, rather than the difference. the ratio should be fairly constant if the 2 methods are measuring the same phenomenon.

c The tabulated critical values of the statistical test used for their test of the fly counts on line 1 of Table I are not easily available for the n's of 40 and 55. How then did the authors determine the significance level? {if you answer "by computer package", then say what formula the computer package used}.

Use the large sample approximation by the Gaussian distribution. If T is the sum of the ranks in the smaller sample, then (as in A&B, as in notes p. , etc.),

$$E(T) = \frac{n_{\text{smaller}}\{n_{\text{smaller}} + n_{\text{bigger}} + 1\}}{2} \quad \text{and}$$

$$\text{Var}(T) = \frac{\{n_{\text{smaller}}\}\{n_{\text{bigger}}\}\{n_{\text{smaller}} + n_{\text{bigger}} + 1\}}{12}$$

d In the 1988(I) control base, the mean of the 55 daily counts was 9.5(SEM 1.5). From this, reconstruct the SD of the 55 counts and say whether it 'fits with' a count distribution one might have considered in an earlier question above. What about the counts for the other sessions and bases? If it doesn't fit, explain why.

$$\text{SEM} = \frac{\text{SD}_{\text{individuals}}}{\sqrt{n}} \quad \text{so that } \text{SD}_{\text{individuals}} = \sqrt{n}\ \text{SEM}.$$

So, in this e.g., SD = $\sqrt{55}$ 1.5 ≈ 11.

If the individual counts followed a homogeneous Poisson distribution with a mean of 9.5, then we would expect their variance to be approximately 9, i.e. their SD to be about 3. In fact there is considerably greater-than Poisson temporal variation.

e# Verify the p=0.024 for the comparison of the mean for all sessions (performed by the conservative approach).

They had 4 paired observations, and so performed a paired t-test on the within-pair differences.

Results -- bacteriological culture.

a* *Shigella* sp was isolated from 2 (6%) of the 33 pooled samples. Use this estimate of 6%, and follow the logic you used in **Materials and methods -- bacterial culture of flies** above, to estimate the % of individual flies from whom shigellae were "isolate–able". Give a 95% confidence interval for your estimate.

If π_{indiv} represents the proportion of positive flies and π_{pool} the proportion of

pooled samples that would be positive, then

$\pi_{pool} = 1 - (1 - \pi_{indiv})^{10}$, so that

$\pi_{indiv} = 1 - (1 - \pi_{pool})^{1/10}$, so that

(*) estimate of $\pi_{indiv} = 1 - (1 - \text{estimate of } \pi_{pool})^{1/10}$ i.e.

estimate of $\pi_{indiv} = 1 - (1 - 2/33)^{1/10} = 0.0062 = 0.62\%$

CI for π_{indiv} can be obtained by substituting the upper and lower limits for π_{pool} [obtained from the binomial proportion 2/33] into (*).

Results -- incidence of diarrhoea and shigellosis.

a# Calculate a p-value for the 12% vs. 17% incidence of diarrhoea in session II in 1988; do the same for session I 1989.

χ^2 test or z-test of 2 independent proportions

b Defend the authors' use of one-tailed (or, more accurately speaking, one-sided) hypotheses.

They find it difficult to imagine that decreasing the fly count would <u>increase</u> shigellosis.

c Were all of the statistical tests in Table I guided by a 1-sided alternative?

If they used χ^2 tests, and looked up $P(\chi^2 \geq \text{observed})$, they are implicitly performing 2-sided tests, since the χ^2 is a squared deviation, and so, unlike the z-test for proportions, does not indicate the direction of the difference.

d Do you agree that *" Intensive fly control resulted in significant reductions in incidence of shigellosis during three of the four training sessions "* ?

It seems that just 2 of the 4 were "significant"

e# Calculate a p-value for the 1.1% vs. 6.7% incidence of shigellosis in session I in 1988.

It can be done using the χ^2 test, since the <u>expected</u> numbers are > 5.

f Why is the rate of seroconversion for Shigella O antibodies the *most objective measure* of transmission of shigellae? {end of 4th paragraph of Discussion}.

If soldiers, or those recording the process and the outcomes, were not blinded to the intervention, one could imagine the possibility of less-than-objective recordings. Also, the author's definition of shigellosis may have be subject to sampling problems (this would produce under-estimates, but would presumably affect both sides of the comparison equally). The definition of seroconversion is more numerical.

Percentage reduction is used throughout as the comparative parameter to measure efficacy. The confidence intervals for this are reported only when the % reduction is based on proportions, presumably using formulae for the SE(%reduction) or SE(log[%reduction]) found in such books as "KKM" or Walker. The authors could have used formula 3.15 in A&B to calculate CI's for %reductions in means.

g# What if the parameter of interest were the absolute reduction in mean fly count? For any one session (i.e. one row of table I) , how would one calculate CI's for these absolute differences? Do you feel comfortable calculating a CI for the absolute reduction in disease incidence in any one row?

CI for the Δ of 2 means. The SE of the difference in means can be calculated using the SEM's calculated from the separate variances in the 2 sessions; the n's are large and so the t or z table will be fairly accurate.

It is the same idea, with proportions rather than means. The only question is whether the n soldiers in the base can be treated as n independent observations (do some of the episodes have a common source?) That is why it is so important to have the trial repeated 4 times.

Results -- additional analyses.

One gets the impression that these were added at the insistence of a reviewer, but that the authors did not really agree with this second approach (i.e. the aanalyses sound like they were added to avoid alienating the reviewers rather than "for the sake of completeness").

a# Carry out a "conservative" test to compare the mean fly counts according to the logic used in the data presentation in Table II

Paired t-test (3 df) on the 4 within-row differences

b What other "conservative" test(s) is(are) available? That null/alternative hypotheses do they test?

Sign Test or Wilcoxon Signed Rank Test. They test whether the median difference is = 0 vs. > 0.

Discussion -- 4th paragraph.

a# Do the "two discordant findings" give you additional arguments for considering the session-base, rather than the soldier, as the unit of analysis? Explain.

Yes, they show that there are many other communal influences. These influences not going to be minimized by having a bigger n per base, but rather by having a greater number of bases.

b Do you agree with the authors' conclusions?.

Generally, yes. I find the study fairly convincing. The strength comes from the 4 comparisons rather than from the impressive p-values in any one base in any one year.

Abstract and Figure 2 of Article: Compensatory enlargement of human atherosclerotic coronary arteries NEJM 1987; 316:1371-5

Abstract

a What does the "$P \leq 0.001$" following the r=0.44 mean in words?

If the 2 variables were uncorrelated, and if all that was at work was random (sampling) variation, then the probability of obtaining $|r| \geq 0.44$ is less than 0.1%.

b# What does the "0.88(lesion area)" in the regression mean in words?

the average y in those with a lesion area of x+1 is on average 0.88 higher than it is in those with a lesion area of x.

Figure 2

a Give <u>your</u> interpretation of the r=0.44. Do so by completing the following sentence : "The correlation was significant (r=0.44, P<0.001) indicating that ... "

It is unlikely that the correlation ρ in this class of patients is zero. Just because the p-value for testing $\rho=0$ is impressive doesn't necessarily mean that ρ is very strong (the p-value is affected by the large n).

b# In 10 ≤ number of words ≤ 30, complete the sentence: "In Figure 2, the unbroken line represents ... "

an estimate of a straight line joining the series of means -- one mean y for every value along the x range in question.

c# How do you think the authors calculated the broken lines? Would you do otherwise? Explain what purpose you want the lines to serve.

I think the 4.8 is the √ of the average squared residual i.e. the 4.8 is the standard deviation of the residuals. This would mean that, if the y values at each x had a Gaussian distribution around the line, ≈ 68% would be within 1 SD=4.8 of the line. They seems to have drawn the lines:
intercept + slope•X ± 1 . SD_{res} . $\sqrt{1}$.

So it looks like they tried to make a 68% range for individual values; unfortunately, they neglected to add in the statistical uncertainty concerning the location of the line -- had they done so the range would have been wider with bow-shaped boundaries. The proper formula for a 68% range of individuals at any X is given by

$$\text{intercept} + \text{slope}\bullet X \pm t_{df} \cdot SD_{res} \cdot \sqrt{1 + \frac{1}{n} + \frac{[X-\text{xbar}]^2}{\sum[x-\text{xbar}]^2}} \ .$$

Since the df are large, the t-value which encompasses 68% of observations is close to the corresponding z-value, namely z=1. A 95% range is obtained by using z=1.96.

Note that these ranges will only be accurate at all x if (i) the line is an adequate fit (ii) the vertical variation around the line is Gaussian and (iii) the same size at all values of x.