Hypothesis Testing *vs.* Statistical Estimation Procedures in Caries Clinical Trials

J. W. FERTIG

**Emeritus, Division of Biostatistics, Columbia University School of Public Health, New York, New York*

J Dent Res 63(Spec Iss):808-811, May, 1984

Introduction.

In the usual caries clinical trial, we have two measurements of the caries status of the mouth: a measurement prior to the initiation of the study (x_1), and a measurement after several years (x_2). Intermediate measurements are often made but are ignored here. A frequent measurement of the caries status is the DMFS index – the number of decayed, missing, and filled surfaces. The final measurement, x_2, is correlated with the initial value, x_1. The usual way of taking this relation into consideration – *i.e.*, of adjusting x_2 – is to use the increment $y = (x_2 - x_1)$, which is positive except for reversals. This method of adjustment makes the tacit assumption that x_2 is linearly related to x_1, with a slope of unity, or, equivalently, that y is linearly related to x_1, with a slope of zero. When the assumption of linearity but not that of slope is met, we can employ the analysis of covariance technique, in which we fit the correct relation and adjust accordingly. For our present purpose, we assume that the use of the increment (y) as the outcome variable is satisfactory.

In the usual clinical trial, a group of available children is divided at random into two (or more) groups, one to receive a presumably active treatment (t), the other a control treatment (c). The control treatment is often a placebo, which should be just like the presumably active treatment (toothpaste, mouthwash, etc.), except that it does not contain the active ingredient. Furthermore, the study is to be double-blind – *i.e.*, neither the examiner nor the subject knows what treatment the subject receives. In a well-conceived study, the number of subjects in the control groups, N_c, and the number in the treated group, N_t, are equal or nearly so. Unfortunately, in the course of the study, drop-outs from one or both groups may occur, so that the remaining numbers, n_c and n_t, which we have to use in a comparison may differ to some extent. The inequality is not as disturbing as the fact that the drop-outs may be selective for caries status, and the selectivity may differ among the groups. Pre-stratification of the subjects by the baseline value, x_1, does not automatically take care of the drop-out problem.

Comparison of two treatments.

The distribution of the increment, y, is not a normal curve. The variable is discrete. Furthermore, the distribution is usually positively skewed. This is often the case with distributions where there is a fixed lower limit which is frequently attained or approached, but no upper limit. Nevertheless, the mean is a descriptive constant. We want to compare the mean increments, $\bar{y}_c$ and $\bar{y}_t$, of the two groups.

In view of the usual sample sizes, n_c and n_t, in clinical caries studies, often 50 or more, we may assume that the sampling distribution of means $\bar{y}_c$ and $\bar{y}_t$ is essentially normal. In fact, we are concerned with the sampling distribution of the difference $(\bar{y}_c - \bar{y}_t)$, and that distribution will be closer to normal than those of $\bar{y}_c$ and $\bar{y}_t$ separately. Therefore, we have no hesitation in using statistical procedures based on normal curve theory. This ordinarily implies the use of the *t* test as the test of significance.

Test of significance.

To use the *t* test, the variances s_c^2 and s_t^2 of the two groups are pooled to give:

$$s^2 = [(n_c-1)s_c^2 + (n_t-1)s_t^2] \,/\, (n_c + n_t - 2).$$

The estimated standard error (SE) of the difference in mean increments is:

$$SE(\bar{y}_c - \bar{y}_t) = \sqrt{\frac{s^2}{n_c} + \frac{s^2}{n_t}}\,.$$

Then the ratio

$$t = (\bar{y}_c - \bar{y}_t)/SE$$

is taken into the *t* distribution, with degrees of freedom (df) equal to $(n_c + n_t - 2)$. With the number of df available, the critical value of *t* is practically the same as that of the normal curve – say, 1.96 at the 5% (two-tailed or two-sided) level of significance.

The assumption of equality of the true variances (homoscedasticity) may be valid when the two treatments are the same as under the null hypothesis in a significance test, or nearly the same. However, if the treatments differ greatly, the distribution with the greater mean often has the greater variance. This is rather typical of highly skewed distributions. Many workers feel more comfortable about pooling when testing a null hypothesis (significance test) than when estimating the true difference (confidence interval). The practice of testing the homoscedasticity before pooling is usually frowned upon because it may change the significance level of the significance test of the means in some unknown manner. The matter of pooling is almost academic, since, when $n_c = n_t$, the same SE is obtained whether one pools or not.

If we do not pool the variances, our estimate of the SE is

$$SE(\bar{y}_c - \bar{y}_t) = \sqrt{\frac{s_c^2}{n_c} + \frac{s_t^2}{n_t}}\,.$$

Theoretically, the ratio of $(\bar{y}_c - \bar{y}_t)$ to this SE is not distributed as *t*, and one may not feel comfortable using the ratio as *t*. One can always justify referring the ratio to the normal distribution, since the sample variances are based on large values of n_c and n_t and are thus close estimates of the true variances. In the Table, we compare the means of the control group with that of the treated in one of the studies cited by Dubey *et al.*[1]

One-sided *vs.* two-sided tests.

The investigator wishes to test the null hypothesis, H_o, that $\mu_c = \mu_t$ or that $(\mu_c - \mu_t) = 0$ at some level of significance –

*Present address, 5 Ironwood Lane, Woodbury, CT 07798

TABLE
VARIOUS COMPARISONS OF MEAN DMFS INCREMENTS OF A TREATED AND A CONTROL (PLACEBO) GROUP

Control	n_c=215, $\overline{y}_c = 4.49$, s_c^2=20.16, pooled s^2=16.66
Treated	n_t=190, $\overline{y}_t = 3.57$, s_t^2=12.70
	$(\overline{y}_c - \overline{y}_t) = 0.92$; 95% CI=0.13 to 1.71
	$SE(\overline{y}_c - \overline{y}_t) = 0.401(0.406)$*
	t = 2.29(2.27)*; P=0.022(0.023)*
	% reduction = 20.5%; CI=3.0% to 35.6%[6]

*Values in parenthesis refer to use of separate variances.

say, α=5% – where α is the probability of rejecting H_o when it is true. If the question is whether *t* is better than c (smaller increments), then evidently suitable alternatives to H_o are that $\mu_c > \mu_t$ or $(\mu_c - \mu_t) > 0$. In fact, by so limiting the alternatives, the null hypothesis is re-stated as $(\mu_c - \mu_t) \leq 0$. In that case, all of α is put in the upper tail of the sampling distribution centered at zero ($\overline{y}_c - \overline{y}_t$ large positive). Evidently, the investigator is willing to state that the case where $(\overline{y}_c - \overline{y}_t) < 0$ cannot happen, does not interest him, or is due to chance. But the case $(y_c - y_t) < 0$ does happen and often does interest the investigator who wants to test it. His statement of the null hypothesis and alternatives does not permit him to test the difference. Some investigators inadvertently decide on the alternatives (which tail to use) after they see their difference. This is not legitimate unless they agree to use just half of the sampling distribution, in which case the test is equivalent to a two-sided test.

In a two-sided test, the alternatives to the null hypothesis are $(\mu_c - \mu_t) \neq 0$, which permits one to test negative as well as positive differences. The level of significance, α, is divided, usually equally, among the two tails, *i.e.*, $\alpha/2$ in each tail. The probability P of exceeding the difference in either direction is, of course, twice that of the one-sided test.

Armitage[2] (p. 104) puts the case for a two-sided test well: "Before the data are examined, one should decide to use a one-sided test only if it is quite certain that departures in one particular direction will always be ascribed to chance, and therefore regarded as nonsignificant, however large they are. This situation rarely arises in practice, and it will be safe to assume that significance tests would almost always be two-sided."

Confidence interval (CI).

Often the difference $(\overline{y}_c - \overline{y}_t)$ is large but not statistically significant at the α level, in spite of the rather large values of n_c and n_t, because of the large inter-individual variation of the increment, s_c^2 and s_t^2. This does not mean that the study is necessarily negative and that we should be discouraged. In fact, if we compute the $(1-\alpha)$ confidence interval (CI), we find the whole array of possible true values of $(\mu_c - \mu_t)$ which are compatible with the data, *i.e.*, are not rejected by a two-sided significance test at the α level. Of course, we may also construct the CI when $(\overline{y}_c - \overline{y}_t)$ is significant. In that case, the interval does not include zero, but lies to one side or the other of zero.

If we operate at the 5% significance level, we should use the 95% CI given by $(\overline{y}_c - \overline{y}_t) \pm 1.96$ SE or, more elegantly, by

$$(\overline{y}_c - \overline{y}_t) - 1.96\ SE \leq (\mu_c - \mu_t) \leq (\overline{y}_c - \overline{y}_t) + 1.96\ SE.$$

95% of the time, such intervals will cover the indicated parameter, but, in a given instance, the interval either does or does not cover the parameter.

Let $(\overline{y}_c - \overline{y}_t)$ be positive and not significant. Then the lower confidence limit is negative and the upper limit is positive. $(\overline{y}_c - \overline{y}_t)$ is then compatible with a true difference of zero as well as with a true positive difference as large as $(\overline{y}_c - \overline{y}_t) + 1.96$ SE, the upper confidence limit, or a difference as small as $(\overline{y}_c - \overline{y}_t) - 1.96$ SE, the lower confidence limit.

We see from the Table that the CI is practically the same whether we pool or not when n_c and n_t are equal or practically so. The CI agrees with the two-sided significance test, in the sense that if the difference was significant at the α level, then zero is not included in the $(1-\alpha)$ CI, and if the difference was not significant at the α level, then zero is included in the CI. This agreement between the two-sided significance test and the two-sided CI is not complete if one pools the variances for the significance test but not for the CI.

Obviously, if a one-sided significance test is appropriate, then a one-sided or asymmetrical CI is appropriate. For the case where the alternatives to the null hypothesis are that $(\mu_c - \mu_t) > 0$, the CI extends from $(\overline{y}_c - \overline{y}_t) - 1.64$ SE to infinity for $(\overline{y}_c - \overline{y}_t)$ positive. The use of such asymmetrical CI is rather unusual.

Negative studies and equivalent treatments.

A negative study might be defined as one in which $(\overline{y}_c - \overline{y}_t)$ does not differ significantly from zero and in which neither the upper nor lower confidence limit reaches any meaningful value of $(\mu_c - \mu_t)$ – say, Δ – as defined by clinical considerations. Thus, $(\mu_c - \mu_t)$ may well differ from zero, but the difference is of no practical value. Under these circumstances, the two treatments may be said to be "equivalent". Complete equivalence, of course, implies that $(\mu_c - \mu_t)$=0, so that zero should be a value compatible with the observations.

Considerations of sample size and power for significance tests.

In our discussions thus far, we have focused on the data which are available from some study of reasonable size. We have not considered whether that size is large enough to discriminate between competing parameter values $(\mu_c - \mu_t)$ nor large enough to narrow the CI adequately. We have been entirely concerned with the probability of the first type of error α of falsely rejecting a true null hypothesis or the confidence coefficient $(1-\alpha)$, the long-run proportion of CI's that will contain their parameter.

When $(\overline{y}_c - \overline{y}_t)$ is not significant but the CI is wide, so that the limits are much larger in absolute value than some meaningful value, Δ, the study lacks sufficient sensitivity or power to distinguish between zero and Δ, either because the sample sizes are too small or because the variance of the increment (s_c^2, s_t^2) is too large. In planning a study with sufficient power to distinguish between alternatives, we need to consider also the second type of error, β, the probability of not rejecting the null hypothesis when in fact Δ is the true value of $(\mu_c - \mu_t)$, or equivalently the power $(1-\beta)$, the probability of rejecting the null hypothesis when Δ is the correct value (Snedecor and Cochran[3]; Chilton[4]). The usual choice of α (two-sided) is 5%, corresponding to a critical value z_1 on the normal curve of 1.96. Choices of the power $(1-\beta)$ of 80%, 95%, and 97.5%

correspond to critical values z_2 of 0.84, 1.64, and 1.96, respectively. It should be noted that all of β is put into one tail of the normal curve centered at Δ and facing zero.

Appropriate information on the variance of the increment, s^2, is obtained from past studies or other available information. Thus, for equal sample sizes and pooling, SE is $\sqrt{2s^2/n}$. To accomplish the objective, n for each group must be sufficiently large so that Δ corresponds to (z_1+z_2) SE. This gives a solution

$$n = 2s^2(z_1+z_2)^2/\Delta^2.$$

With this value of n, non-significance for the observed $(\bar{y}_c-\bar{y}_t)$ implies that the absolute value of $(\mu_c-\mu_t)<\Delta$, although it may actually be greater than zero. Significance implies that $(\mu_c-\mu_t)\neq 0$. We are not really "detecting Δ". The fact that our estimate of n is imprecise because of having imprecise values of s^2 need not concern us overly, since at any rate we are going to do a significance test with the actual data after the study is completed.

Values of s seen in many studies range around 4. With this value of s and assuming $\Delta=1$, we compute n for $\alpha=0.05$, $\beta=0.20$ (power=80%).

$$n = 2(4^2)(1.96 + 0.84)^2\ /1^2=251$$

Many studies have sample sizes of this order of magnitude or larger. We note that, if $\mu_c=5$, Δ of 1 corresponds to a percentage change (reduction) of 20%. For $\beta=0.05$ and 0.025, n=415 and 492, respectively, for each group.

Large sample sizes are obviously required if one wants to ensure discrimination between zero and Δ. If one asks for the required n to make an observed value significant at the 5% level, *i.e.*, if one assumes that by chance $(\bar{y}_c-\bar{y}_t)=\Delta$, then n=123. This corresponds to putting $z_2=0$ and accepting a power of only 50%.

Whether or not n was determined from these power considerations, after the study is completed, one may engage in the exercise of calculating how much power one had against various alternatives. With n as calculated from these power considerations, our power for the alternative Δ will obviously be $(1-\beta)$, except for a possibly bad choice of s.

Sample size considerations for confidence intervals.

We want to determine n for each group so that the upper (and lower) $(1-\alpha)$ confidence limits will not exceed Δ in absolute value and so that zero is contained in the interval. It may be verified that for a symmetrical CI, this implies that $(\bar{y}_c-\bar{y}_t)$ lies between $-\Delta/2$ and $\Delta/2$. Furthermore, $\Delta/2$ corresponds to $z_1\sqrt{2s^2/n}$, where z_1 is the critical value of the normal curve cutting off $\alpha/2$ in each tail, 1.96 for a 95% CI. Solving for n:

$$n = 8(1.96)^2s^2/\Delta^2$$

With this n and $(\bar{y}_c-\bar{y}_t)=\Delta/2$, the upper confidence limit is at Δ, and similarly, if $(\bar{y}_c-\bar{y}_t)= -\Delta/2$, the lower limit is at $-\Delta$. If $(\bar{y}_c-\bar{y}_t)$ lies between $-\Delta/2$ and $\Delta/2$, the absolute values of the upper and lower limits are less than Δ.

With $\Delta=1$, s=4, $(1-\alpha)=0.95$, n is 492. But this is exactly the sample size required for a significance test at the 5% level with a power of 97.5%, which is ordinarily more power than that required in significance testing. It will be recognized that the formula for n is the same as that already given for the significance test if z_2 is put equal to z_1. A table of values of n appropriate for significance tests with various powers will also serve for the CI problem if appropriate asymmetrical CI's are used. For example, the sample size for the significance test at the 5% level with 80% power corresponds to using an asymmetrical 77.5% CI. For $(\bar{y}_c-\bar{y}_t)>0$,

$$(\bar{y}_c-\bar{y}_t) - 1.96\ \text{SE} \leq (\mu_c-\mu_t) \leq (\bar{y}_c-\bar{y}_t) + 0.84\text{SE}.$$

Large sample sizes are obviously required to show equivalence by our definition. If one assumes that by chance the observed difference is exactly zero, then the upper limit of the $(1-\alpha)$ CI is Δ for values of n only one-fourth as large as those obtained by the more restrictive formula, 123 in the example given instead of 492. One can also reduce the value of n by using smaller values of the confidence coefficient $(1-\alpha)$. For $(1-\alpha)=90\%$, n=344; for $(1-\alpha)=80\%$, n=210.

The case of positive controls.

In many areas of dental research in developed countries, placebos are no longer used, so that testing effectiveness of an agent *vs.* a placebo is no longer an option. Rather, a positive control is used, *i.e.*, an agent known to be effective against a placebo. To test whether the agent is better than a positive control, a one-sided significance test may be used. However, a new product should not necessarily have to be better than a proven product. There is room for more than one good product. The question as to whether the test agent differs from a positive control can be handled by a two-sided significance test. In addition, one needs a statement of what difference is unimportant, or what difference is permissible and still consonant with the idea of "equivalent" treatments. Then the significance test is supplemented by a CI. If the CI does not include zero (significant), obviously the two agents are not equivalent. If the CI includes zero (non-significance), and the upper limit is less than Δ, the two agents are said to be equivalent for our purposes, because whatever $(\mu_c-\mu_t)$ may be, it is less than Δ with confidence $(1-\alpha)$. In many studies of limited size, we do not have significance, but the upper limit exceeds Δ, so that we do not have enough power for purposes of establishing "equivalence".

Obviously, in planning such studies to show equivalence of two test agents, the considerations of sample size based on some suitable value of s, Δ, α, and β should enter. The value of n giving sufficient power for a significance test may not be good enough. The upper $(1-\alpha)$ confidence limit when non-significance is found may still exceed Δ, unless of course $\beta=\frac{1}{2}\alpha$, in which case the necessary value of n given by the significance test approach is the same as that given by the CI approach.

Percentage reductions.

Often the difference between two treatments is expressed in relative form as a proportion of the control

$$p = (\bar{y}_c-\bar{y}_t)/\bar{y}_c.$$

In a study with the data at hand, the test of significance of p *vs.* zero has already been made when $(\bar{y}_c-\bar{y}_t)$ was tested against zero. The CI for p can be handled by an application of Fieller's formula for a ratio of random variables (Finney[5], Wallenstein *et al.*[6]), or by an approximation to the standard error (Dubey *et al.*[1]) which is valid when the coefficient of variation of $\bar{y}_c$, $\text{SE}(\bar{y}_c)/\bar{y}_c$, is negligible.

For the 95% CI (Wallenstein *et al.*),

$$\text{CI} = 1 - \frac{1}{1-g}\left(\frac{\bar{y}_t}{\bar{y}_c} \pm \frac{1.96}{\bar{y}_c}\sqrt{\frac{(1-g)s_t^2}{n_t} + \left(\frac{\bar{y}_t}{\bar{y}_c}\right)^2\left(\frac{s_c^2}{n_c}\right)}\right)$$

$$\text{where } g = \frac{(1.96)^2\, s_c^2}{\bar{y}^2 n_c}$$

$$\text{(Dubey } et\ al.\text{)} \quad CI = 1 - \left(\frac{\bar{y}_t}{\bar{y}_c} \pm \frac{1.96}{\bar{y}_c}\sqrt{\frac{s_t^2}{n_t} + \left(\frac{\bar{y}_t}{\bar{y}_c}\right)^2 \left(\frac{s_c^2}{n_c}\right)}\right)$$

s_c^2 and s_t^2 are often pooled.

The definition of equivalence is often stated in terms of a relative reduction as $(\mu_c-\mu_t)/\mu_c={}^{\Delta}/\mu_c$. For calculating the sample size necessary for the significance test, or for the CI to have the desired discriminatory properties for alternatives to the null hypothesis as large as ${}^{\Delta}/\mu_c$, the approximate value of the SE of p may be used. However, it is more convenient to translate the definition of equivalence into terms of Δ by using some appropriate value of μ_c, and then use the corresponding formulae for sample size already given.

Increments adjusted by analysis of covariance.

Frequently, we are not satisfied that the increment $y=(x_2-x_1)$ takes the value of x_2 adequately into account, since the slope of the regression of y on x_1 (x_2 on x_1) is not zero (one). If the regression is linear, we may proceed to the typical analysis of covariance procedure of fitting the regression of y (or x_2) on x_1, making the adjustments to y (or x_2) by the appropriate line, and calculating the significance test and the CI.

To apply the analysis of covariance correctly, several assumptions are made other than normality: linearity: constancy of residual variation around the line for all values of x_1; equality of residual variation of the two groups; and parallelism of the regression lines (equal slopes) for the two groups. The variance around the regression line is probably not constant. The equality of the residual variation for the two groups is questionable, but for large groups of roughly equal sample size this is not a major difficulty. The most important assumption is that of parallelism, or equality of the two slopes estimated by b_c and b_t. For treatments that do not differ from each other, as under the null hypothesis, or for treatments that should not differ greatly, as with positive controls, the assumption of parallelism may be reasonable. Of course, one or more of the assumptions may be tested, but then the level of significance in the final comparison of the adjusted means may be disturbed in some unknown way. When the analysis of the data is made in terms of increments, these problems (which are raised in the analysis of covariance) may also be present, but they are hidden.

If one admits parallelism together with the other assumptions, one proceeds with the typical analysis of covariance, including the CI. All the problems previously discussed in connection with significance tests and CI's apply. With sufficient background information, one may calculate sample sizes necessary to discriminate satisfactorily between alternative hypotheses. The calculation of CI for the relative difference between adjusted $\bar{y}_c$ and $\bar{y}_t$ is more complicated than for the case of the unadjusted difference (Wallenstein *et al.*[6]).

If we do not admit parallelism, there is no unique adjusted $(\bar{y}_c-\bar{y}_t)$. One may compute adjusted $\bar{y}_c$ and $\bar{y}_t$ and hence the difference at one or more values of x_1 — in particular, at the overall mean $\bar{x}_1$.

$$\text{adj } \bar{y}_c=\bar{y}_c-b_c(\bar{x}_c-\bar{x}_1)$$
$$\text{adj } \bar{y}_t=\bar{y}_t-b_t(\bar{x}_t-\bar{x}_1)$$

This difference may be tested for significance (Chilton[4]). The comparison of two such adjusted means corresponds closely to the test of the treatment effect in a two-way layout analysis of variance when there is an interaction of treatment with strata based on the value of x_1.

Conclusions.

Significance tests and confidence intervals have been discussed for clinical trials. The two approaches are complementary, being based on the same background theory. The confidence interval approach is often the more informative, especially when the control group is a positive control rather than a placebo, and equivalence of a new agent with the control is to be established. Sample size considerations are discussed for both the significance test and the confidence interval.

REFERENCES

1. DUBEY, S.D.; LEHNHOFF, R.W.; and RADIKE, A.W.: A Statistical Confidence Interval for the True Percent Reduction in Caries Incidence Studies, *J Dent Res* 44:921-923, 1965.
2. ARMITAGE, P.: **Statistical Methods in Medical Research.** New York: John Wiley and Sons, 1971.
3. SNEDECOR, G. W. and COCHRAN, W. G.: **Statistical Methods,** 7th ed., Ames, Iowa: The Iowa State University Press, 1980.
4. CHILTON, N.W.: **Design and Analysis in Dental and Oral Research,** 2nd ed., New York: Praeger Publishers, CBS, Inc., 1982.
5. FINNEY, D.J.: **Statistical Methods in Biological Assay,** 3rd ed., London: Charles Griffin & Co., 1978.
6. WALLENSTEIN, S.: FLEISS, J.L.; and CHILTON, N.W.: Confidence Intervals for Percentage Reduction in Caries Increments, *J Dent Res* 61:828-830, 1982.