SAMPLING

In STI interventions, size matters

R A Crosby, R Rothenberg

The investigator must juggle sample size and effect size to produce a study with a priori credibility and ex post facto utility

Sample size and effect size are interrelated parameters that have been given insufficient consideration in analyses of two major outcome measures in the field of sexually transmitted infections (STI) prevention—STI incidence and condom use. Their interrelation highlights two critical features of interventions—statistical significance and epidemiological importance.

Though there are a myriad variations on the theme of sample size, the calculation usually depends on four parameters. The investigator must designate an acceptable level for type I error (the probability of falsely rejecting a null hypothesis); type II error (the probability of failing to reject a false null hypothesis); variance (the amount of dispersion in the result that would be acceptable); and the effect size (the size of the detectable difference that is deemed important). Cynics will be quick to point out that sample size depends on only one parameter—the amount of money available for the study—but even if true, researchers must still deal with the consequences of the terms so dictated. Because type I and type II error are often set by convention, the investigator must juggle sample size and effect size to produce a study with a priori credibility and ex post facto utility.

SOME STATISTICAL CONSIDERATIONS

An important difference among studies is whether or not they are measuring continuous or dichotomous outcomes. In the former case, effect size can be assessed by transformation into units of standard error, and a test of statistical significance is straightforward. Yet, continuous outcomes are rare in STI research. Incidence of STIs is inherently a dichotomous outcome and condom use is typically dichotomised, based on the assumption that consistent (that is, 100%) use is protective against infection whereas lower rates of use are not protective.[1–3]

For dichotomous outcomes, the investigator measures a proportion (for example, frequency of condom use or the frequency of STIs) in a treatment and control group. The difference of proportions (the effect size) is usually tested for statistical significance with the same transformation into units of standard error (the normal approximation), without realising that such a transformation is critically dependent on the underlying frequency of the event in the population. The variance of a proportion differs along the continuum from zero to one (that is, the variance is non-uniform or *heteroscedastic*) and thus the same effect size will have different significance depending on where the outcome, p, falls across the zero to one spectrum (fig 1 (z test, probability)).

Though in practice little attention has been paid to heteroscedasticity, some important statistical work has focused on stabilising the variance of the binomial distribution (see appendix). The inverse sine (arcsin, or $\sin^{-1}$) transformation, specifically,

$$P = 2\arcsin\sqrt{p}$$

is one of several transformations that modulates the variance for extreme values of p.[4] A test of significance using this transformation demonstrates that, for the case of a 4% difference and a moderately large sample size (n = 400), the boundaries of non-significance are extended from 10% to 3% (fig 1). In other words, a small difference with population probability as high as 10% would be significant without the transformation, but loses significance if a procedure to stabilise the variance is invoked. The ability of the transformation to correct for a changing variance diminishes as the expected value of the outcome decreases and as the sample size increases. Thus, after a certain point, a low expected value and a large sample size can ''overpower'' the transformation and it will no longer render non-significance for a value that was significant before the transformation.

EFFECT SIZE, SAMPLE SIZE, AND POWER

Cohen used this transformation,

$$P = 2\arcsin\sqrt{p}$$

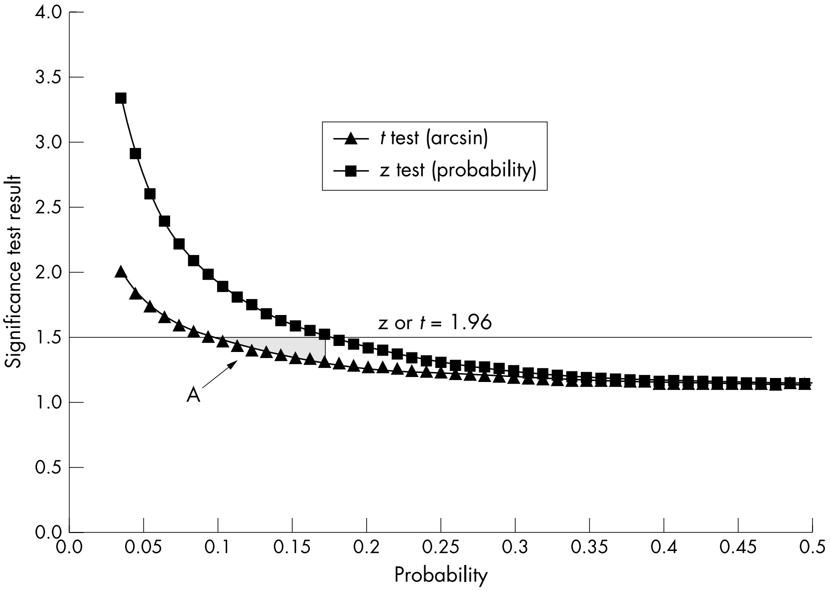

Figure 1 Comparison of statistical tests on a difference of 0.05 detected between proportions with a sample size of 400 in each group, using probabilities directly or the arcsin transformation of the probability*. *Each point on the curve from 0.01 to 0.50 was assumed to be the average probability for treatment and control groups that were symmetric about this mid point (that is, $p \pm 0.025$). The arcsin transformation (arcsin $\sqrt{p}$) produces the angle, in radians, whose sin is $\sqrt{p}$; the pooled variance associated with transformations of $-0.02p$ and $+0.02p$ were used to test the significance of the difference in radians (see appendix). Significance testing for the untransformed probability used the typical critical ratio for a difference of proportions $[(p_2 - p_1)/\sqrt{2pq/n}]$. ''A'' represents the area of significance for untransformed test and nonsignicance for transformed tests (see text).

Table 1 Selected (untransformed) proportions necessary to achieve provided h values in studies using STI incidence as the dependent variable

Small effects (h≅0.20)		Medium effects (h≅0.50)		Large effects (h≅0.80)	
STI proportions (%)	h Value	STI proportions (%)	h Value	STI proportions (%)	h Value
2 *v* 6	0.21	2 *v* 15	0.51	2 *v* 27	0.81
3 *v* 7	0.19	3 *v* 17	0.50	3 *v* 30	0.81
4 *v* 9	0.21	4 *v* 19	50	4 *v* 32	0.80
5 *v* 11	0.19	5 *v* 21	0.50	5 *v* 35	0.81
6 *v* 12	0.21	6 *v* 23	0.50	6 *v* 37	0.81
7 *v* 13	0.20	7 *v* 25	0.51	7 *v* 39	0.81
8 *v* 14	0.19	8 *v* 26	0.50	8 *v* 40	0.80
9 *v* 16	0.21	9 *v* 28	0.50	9 *v* 42	0.80
10 *v* 17	0.17	10 *v* 30	0.51	10 *v* 44	0.81

to form the basis of the measurement h—the difference of transformed proportions.[5] In an in-depth discussion, he provided extensive tables that permit calculation of sample size and power based on the transformed variable. He suggested that h values of 0.20, 0.50, and 0.80 constitute small, medium, and large effect sizes, respectively. Table 1 displays selected untransformed proportions that would be necessary to achieve each of these h values in studies using STI incidence as the dependent variable. Proportions listed under the heading ''small effects'' (using the h criteria) would yield risk ratios (in intervention studies) ranging from 0.33 (for a comparison of 2% incidence in the treatment group *v* 6% in the control group) to 0.58 (for a comparison of 10% incidence in the treatment group *v* 17% in the control group). Similarly, under the column showing medium effect sizes, corresponding relative risk ratios range from 0.13 to 0.33. Under the column showing large effect sizes, the range is from 0.07 to 0.23.

Table 2 displays selected untransformed proportions that would be necessary to achieve these same h values in studies with condom use as the dependent variable. Proportions of people using condoms consistently in each of two groups (for example, intervention *v* control) are shown. Differences of proportions that correspond to small effect sizes (h≈0.20) would yield relative risks of 0.73 (first row entry) to 0.88 (final row entry). Differences of proportions corresponding to medium effect sizes (h≈0,50) would yield relative risk ratios of 0.51 (first row entry) to 0.67 (final row entry). Finally, the large effects (h≈0.80) would yield relative risks of 0.39 (first row entry) to 0.58 (final row entry).

Using these correspondences, the h value provides an alternative approach to assessing power in a study. Its advantage lies in making the assessment of effect size less dependent on the underlying population proportion. To achieve 80% power ($\alpha = 0.05$, two tailed) to detect an h value of 0.40, a study would need a harmonic mean of 100 participants (fig 2). The harmonic mean

$$\left[2\frac{n_1 n_2}{n_1 + n_2}\right]$$

is used to account for the likely lack of equal numbers in the two groups. It is especially important in observational studies wherein groups are naturally formed (for example, comparing STI positive and STI negative participants). Such a study would be able to detect a difference of 12% to 19% over a substantial portion of the zero to one range. To achieve 80% power to detect an h value of 0.30, a study would need a harmonic mean of 180 participants; for an h value of 0.20, 400 participants would be required. In the latter case, a difference of 5% to 10% would be detectable over most of the proportional range. Finally, an h value of 0.10 requires about 3000 participants and would detect a proportional difference of less than 5% over most of the range.

QUESTIONABLE SIGNIFICANCE

Several recent STI interventions illustrate the relations described here (table 3). Ford *et al*[6] reported a significant decrease in syphilis in response to an intervention project directed to female sex workers in Bali, with an effect size of 3.2% (6.8% *v* 3.6%), an h of 0.06 (quite small by Cohen's criteria), and a moderately large sample size. When the probabilities are transformed, this result is non-significant. On the other hand, the medium h of 0.45 for gonorrhoea (the population probability was centred at p~0.5) produced a highly significant result with either test.

Golden and colleagues[7] reported a decline in gonorrhoea for men who had been exposed to pretest and post-test HIV counselling (gonorrhoea in women actually increased, and the changes in syphilis were non-significant). With a sample size in the range of 400–500, the effect size for gonorrhoea in men was 4.7% (8.3% *v* 3.6%; h = 0.09) and the highly significant result was of only marginal significance using the transformed test. (As a general practice, authors will term a p value of 0.06 ''marginal,'' but rarely do so for the equally marginal p value of 0.04.) Finally, O'Donnell *et al*[8] reported an effect size of 4.3% (26.8% *v* 22.5%;

Table 2 Selected (untransformed) proportions necessary to achieve provided h values in studies designating condom use as the dependent variable

Small effects (h≅0.20)		Medium effects (h≅0.50)		Large effects (h≅0.80)	
STI proportions (%)	h Value	STI proportions %)	h Value	STI proportions (%)	h Value
25 *v* 34	0.20	25 *v* 49	0.51	25 *v* 64	0.81
30 *v* 40	0.21	30 *v* 54	0.50	30 *v* 69	0.81
35 *v* 45	0.21	35 *v* 60	0.50	35 *v* 74	0.80
40 *v* 50	0.20	40 *v* 65	0.51	40 *v* 78	0.80
45 *v* 55	0.20	45 *v* 70	0.50	45 *v* 82	0.81
50 *v* 60	0.20	50 *v* 75	0.51	50 *v* 86	0.81

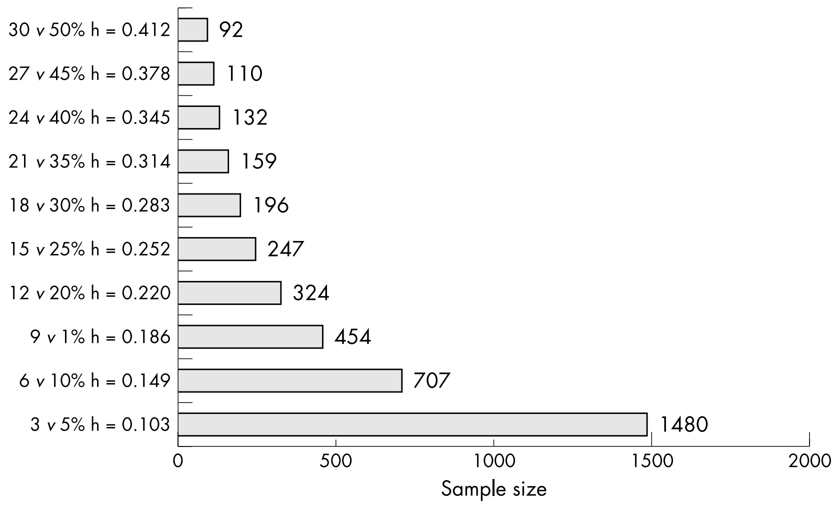

Figure 2 Required number (each of two groups) to achieve 80% power with α set at 0.05, using a two tailed test of significance.

h = 0.09) using a moderately large sample to assess the effect of a video intervention on the occurrence of new STD. Their result, significant without transformation, was marginally significant ($t = 2.02$) after transformation.

LARGE SAMPLE SIZE

The analysis of a CDC study known as Project Respect[9] further highlights the effect of a very large sample. This efficacy trial found that STI incidence (assessed 6 months after study enrolment) among those assigned to an enhanced counselling condition was significantly (p<0.001) lower than incidence among those in the control condition. The values for incidence were 7.2% versus 10.4%, for an effect size of 3.2% (h = 0.11, a very small value by Cohen's criteria). The investigators used a sample size almost four times larger (approximately 1450 in each group) than that depicted in figure 1. As a result, both transformed and untransformed approaches are significant (sample size overwhelms the effect of variance stabilisation). Viewed another way, at a population prevalence of 8%, an effect size of 3% will begin to be significant for both transformed ($t = 2.00$) and untransformed (z = 2.71) tests with a sample size of 1200. However, with a sample size of 1450, the study could have detected an effect size as small as 2% by using the untransformed test.

These observations suggest that there may be important information in the area between significance as calculated for transformed and untransformed probabilities (area ''A'' in fig 1). If we assume that area ''A'' represents a type I error (incorrectly rejecting the null hypothesis), such an occurrence suggests that a large sample size may be capable of detecting an effect size that is not meaningful. Sample size calculations, in circumstances of large, expensive studies that are intervening on low probability events, should thus be based on the h value.

SIGNIFICANCE VERSUS EPIDEMIOLOGICAL IMPORTANCE

Significance testing in the absence of effect size estimation begs the question of whether statistitically meaningful findings have any practical epidemiological importance. This observation, however, brings tough questions to the surface. For example, what is the basis for saying that Project Respect's result, a statistically significant difference of 3.2%, is very small, and how do we sort out the influence of context on such a judgment? It may be argued, for example, that the ''small'' difference could represent a substantial number of cases averted if the enhanced counselling protocol were widely applied in STI clinics. On the other hand, an efficacy trial creates optimal conditions for an intervention effect, and the transition from optimal study conditions to routine clinic setting dictates a considerable shortfall in reaching the maximum effect of 3.2%. Such arguments are difficult to resolve, especially if researchers solely rely on significance testing. A similar example can be found in a review of findings from Project Light, a multisite HIV prevention trial.[10] For example, the study found a significant (p<0.05) difference in STI incidence (as assessed by a 1 year chart review) between participants in the control (5.0%) and experimental (3.0%) conditions. After transformation, these proportions yield an h value of 0.10 (a very small effect).

The rigid application of statistical criteria such as a *p* value of 0.05 or less, has been questioned in recent years.[11–15] Proponents of a less rigid approach argue that the *p* value is inappropriate as currently used, and that the point estimate and confidence limits, taken in context, provide readers with considerably more information for making judgments and permit a more coherent approach to judging a body of literature. In a similar vein, qualitative evaluation schemes have been suggested for a number of statistical measures. Cohen's classification of h values into small, medium, and high, is part of a tradition exemplified by the kappa statistic for assessing interobserver variation. Landis and Koch suggested a

Table 3 Effect size assessment and recalculation of statistical significance in selected studies of STI intervention

		Control	Treatment				Untransformed		Transformed	
Study	Outcome	No (%)	No (%)	Effect size (%)	h Value	Cohen's h classification	z test	Result	*t* test	Result
Ford, 2002[6]	Syphilis	6.8 (1489)	3.6 (359)	3.2	0.06	small	2.71	*Significant*	1.73	*NS*
	Trichomoniasis	7.9 (1489)	3.5 (359)	4.4	0.09	small	3.68	Significant	2.36	Significant
	Gonorrhoea	62.0 (1489)	43.0 (359)	19.0	0.45	medium	6.55	Significant	6.51	Significant
	Chlamydia	43.0 (1489)	41.0 (359)	2.0	0.04	small	0.69	NS	0.69	NS
Golden, 1996[7]	Gonorrhoea (men)	8.3 (468)	3.6 (400)	4.7	0.09	small	2.98	*Significant*	2.05	*Marginal*
	Gonorrhoea (women)	2.9 (140)	6.0 (149)	−3.1	−0.06	small	−1.29	NS	−0.83	NS
	Syphilis (men)	1.7 (468)	0.75 (400)	0.95	0.02	small	1.29	NS	0.61	NS
	Syphilis (women)	0.0 (140)	0.67 (149)	−0.67	−0.01	small	−1.00	NS	–	NS
O'Donnell, 1998[8]	New STD	26.8 (794)	22.5 (1210)	4.3	0.09	small	2.17	*Significant*	2.02	*Marginal*

similar three category empirical approach that is now widely in use.[16] The adoption of Cohen's h value schema would, at the very least, provide a benchmark for assessing studies that detect "small" effect sizes. Researchers could then provide a more contextual analysis—including the economic ramifications of a small intervention difference—of the value of putting such findings into practice.

A frequent criticism of studies is that they are underpowered, a phenomenon that is well understood.[17 18] Clearly, studies may be overpowered as well. Detection of a small effect size with a large sample may be flawed on purely statistical grounds, since it may be an artefact of the binomial distribution itself. Clearly, such a result may also be questioned when considering epidemiological importance. The transformation suggested here, coupled with qualitative criteria, may assist investigators in the matter of size—neither too small nor too big.

Sex Transm Infect 2004;**80**:82–85.
doi: 10.1136/sti.2003.007625

Authors' affiliations

R A Crosby, Rollins School of Public Health, Department of Behavioral Sciences and Health Education, and Emory Center for AIDS Research, Atlanta, GA, USA
R Rothenberg, Emory Center for AIDS Research and Emory University School of Medicine, Department of Family and Preventive Medicine, Atlanta, GA, USA

Correspondence to: R Crosby, PhD, Rollins School of Public Health of Emory University, Department of Behavioral Sciences and Health Education, 1518 Clifton Road, NE, Room 542, Atlanta, GA 30322, USA; rcrosby@sph.emory.edu

Accepted for publication 24 November 2003

APPENDIX

In 1950, Freeman and Tukey demonstrated that the transformed probability

$$X = \arcsin\sqrt{\frac{x}{N+1}} + \arcsin\sqrt{\frac{x+1}{N+1}}$$

$$(\text{approximately}, 2\arcsin\sqrt{p})$$

stabilises the variance of the binomial probability.[4] As noted, Cohen used this property to develop the measure h, and provided qualitative guidelines for the size of a difference of proportions (see text).[5] We demonstrated the effect of transformation (fig 1) by examining probabilities (p) in the range of 0.01 to 0.50, and their transformed equivalents in radians

$$2\arcsin\sqrt{p}$$

The variance of p is $p(1 - p)/n$, and the transformed variance was calculated as the first derivative of the transformed variable times the variance of the untransformed variable

$$s^2 = \frac{\sqrt{p}}{2p\sqrt{(1-p)}} \times \frac{p(1-p)}{n}$$

For each probability in the range, and for a given detectable difference and a given sample size (p = 0.04 and n = 400 in fig 1), we calculated the normal deviate using a z test for the untransformed binomial probability

$$(z = \frac{p_2 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

and a t test

$$t_{df=\infty}\frac{2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}}{\sqrt{s_1^2 + s_2^2}}$$

for the transformed variable, where subscripts 1 and 2 correspond to the upper and lower values, symmetrical around p, that furnish the desired difference at a given p. The same approach was applied to determine untransformed and transformed significance in the examples from the literature.

REFERENCES

1 **Crosby RA**, DiClemente RJ, Holtgrave DR, *et al*. Design, measurement, and analytic considerations for testing hypotheses relative to condom effectiveness against non-viral STIs. *Sex Transm Infect* 2002;**78**:228–31.
2 **Crosby RA**. Condom use as a dependent variable: measurement issues relevant to HIV prevention programs. *AIDS Educ Prev* 1998;**10**:448–57.
3 **Ahmed S**, Lutalo T, Wawer M, *et al*. HIV incidence and sexually transmitted disease prevalence associated with condom use: a population study in Raki, Uganda. *AIDS* 2001;**15**:2171–9.
4 **Freeman MF**, Tukey JW. Transformations related to the angular, and the square root. *Ann Mathem Stat* 1950;**21**:607–11.
5 **Cohen J**. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
6 **Ford K**, Wirawan DN, Reed BD, *et al*. The Bali STD/AIDS Study: evaluation of an intervention for sex workers. *Sex Transm Dis* 2002;**21**:50–8.
7 **Golden MR**, Rompalo AM, Fantry L, *et al*. Early intervention for human immunodeficiency virus in Baltimore sexually transmitted diseases clinics: impact on gonorrhea incidence in patients infected with HIV. *Sex Transm Dis* 1996;**23**:370–7.
8 **O'Donnell CR**, O'Donnell L, San Doval A, *et al*. Reductions in STD infections subsequent to an STD clinic visit: using video-based patient education to supplement provider interactions. *Sex Transm Dis* 1998;**25**:161–8.
9 **Kamb ML**, Fishbein M, Douglas JM, *et al*. Efficacy of risk-reduction counseling to prevent human immunodeficiency virus and sexually transmitted diseases: a randomized controlled trial. *JAMA* 1998;**280**:1161–7.
10 **The National Institute of Mental Health Multisite HIV Prevention Trial Group**. The NIMH Multisite HIV Prevention Trial: reducing HIV sexual risk behavior. *Science* 1998;**280**:1889–94.
11 **Browner WS**, Newman TB. Are all significant p-values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987;**257**:2459–63.
12 **Rothman KJ**. Statistics in nonrandomized studies (editorial). *Epidemiology* 1990;**6**:417–18.
13 **Greenland S**. Randomization, statistics, and causl inference. *Epidemiology* 1990;**1**:421–9.
14 **Goodman SN**. p values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993;**137**:485–96.
15 **Feinstein AR**. P-values and confidence intervals: two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998;**51**:355–60.
16 **Landis JR**, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–174.
17 **DerSimonian R**, Charette LJ, McPeek B, *et al*. Reporting on methods in clinical trials. *New Engl J Med* 1982;**306**(22):1332–7.
18 **Weller SC**, A meta-analysis of condom effectiveness in reducing sexually transmitted HIV. *Soc Sci Med* 1993;**36**:1635–44.