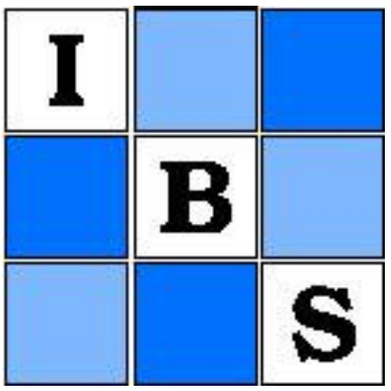

MISCLASSIFICATION IN 2 × 2 TABLES

IRWIN BROSS

Cornell University Medical College

INTRODUCTION

Although in the derivation of the chi-square test (and related techniques) the classifications in the 2 × 2 table are assumed to be correct, there are many practical problems where mistakes in classification are going to be made. The important question then arises: What effects will the misclassification have on conclusions drawn from the usual significance tests? The principal purpose of this paper is to answer this question.

In the medical field there are some classifications that involve almost no risk of error, for example, the categories "lived" and "died". On the other hand, in more complex diagnoses the clinician realizes that there is a considerable risk of error, a risk that may vary a great deal depending on the disease under study, the existence and availability of diagnostic tests, and other factors. For some diseases the principal misclassification will be in the direction of missing some of the actual cases. In other diseases there may be a risk of misdiagnosis in the other direction due to the existence of diseases which "mimic" the characteristics of the disease under study.

Even when more precise classification is possible there may be time or cost factors that necessitate the use of cruder classification methods. In a large survey of mental health, for example, it will rarely be possible to have actual psychiatric diagnoses on all respondents. Instead it will be necessary to classify respondents as "Health Problems" or "Not Health Problems" on the basis of information in a question schedule. One possible procedure in such situations is to construct a mental health scale (using a subsample of "problem" and "non-problem" cases) and thereafter classify individuals by whether they fall above or below some *critical point* on the mental health scale. Such a procedure may well lead to appreciable misclassifications in both directions.

Research workers have quite different opinions as to what can be done in the analysis of data which are subject to misclassifications. Some feel that in a large series the effects will "cancel out" and therefore it is proper to use the usual analysis. At the other extreme are the research workers who argue that the data are biased and therefore no conclusions can be drawn. The findings of this paper represent an intermediate position.

In discussing the effects of misclassification it is important to remember that a given body of data may be suitable for one purpose but not for another. In the mental health survey, for instance, we may want to see if there are different proportions of "problem" cases in two sub-populations (such as ethnic or economic groups) and for this purpose *significance tests* would be used. We may also want to estimate the proportion of "problem" cases in a specified sub-population and for this purpose *confidence intervals* might be used.

Misclassification may present a more serious problem in the case of estimates than in the case of significance tests.

A second point to note is that the performance of a statistical method may be judged by *several* standards. One standard is that a significance test should be *valid.* In other words the risk of a false assertion that "two populations are different" (Type I error) should be some preassigned value such as 5%.

If the same classification system is used in the two populations then misclassification will not affect the validity of the significance test.

Thus, if the experimenter performs chi-square tests in the usual fashion the risk of making false Type I assertions will not be increased by the misclassification. However, there is a price that must be paid for misclassification.

Misclassification tends to reduce the power of the test.

A second standard for judging a significance test is that it should be *powerful.* In other words, it should be able to detect real differences between two populations. Practically speaking the reduction of power corresponds to reduction in effective sample size. Thus in a survey with 2,000 respondents, the misclassification might be equivalent to the "loss" of 500 interviews.

Naturally the extent of loss will depend on the amount of misclassification and a table will be presented which will indicate the magnitudes of the losses in different situations. It is also possible to reduce the loss by taking account of misclassification when the study is planned, for example, by the choice of the *critical point* when a scale is used to make the classifications.

ANALYTICAL MODEL

In order to justify the statements in the introduction and to obtain a more quantitative picture of the effects of misclassification, an analytical model will be set up. As a first step, we will consider effects of misclassification on a sample from a *single* population.

Suppose that a sample of n individuals is drawn from the population. Let u be the *actual* number of "problem" individuals in the sample. Let v be the number of "problem" individuals who are *incorrectly* classified as "non-problem" cases. Let w be the number of "non-problem" individuals who are *incorrectly* classified as "problem" cases. Finally let x be the *apparent* number of "problem" individuals (i.e., the number according to the inaccurate classification system). It follows that $x = u - v + w$.

Now let p be the proportion of "problem" cases in the *population*. Let θ be the probability of misclassifying a "problem" individual. Let ϕ be the probability of misclassifying a "non-problem" individual. Parameters θ and ϕ are characteristics of the *classification system.*

As a result of misclassification the estimated proportion of "problem" cases will not in general be p as before. Instead the expected value of x/n will be:

$$E\left(\frac{x}{n}\right) = p - p\theta + q\phi \qquad \text{where} \qquad q = (1 - p)$$

Although the expectation has been changed, x is a binomial variable and the variance of x obtained* in the usual way, will be:

$$V(x) = n(p - p\theta + q\phi)(q + pe - q\phi)$$

If now:

$$K = 1 - \theta + \frac{q}{p}\phi \tag{1.01}$$

$$K' = (1 - \phi - \theta)^2 + \frac{\theta(1 - \theta)}{q} + \frac{\phi(1 - \phi)}{p} \tag{1.02}$$

then if n is large, the number of "problem" cases, x, will be approximately normally distributed with mean

$$E(x) = npK$$

and variance

$$V(x) = npqK'.$$

If there is no misclassification, $K = K' = 1$.

*This simplified derivation was suggested by Referee I.

If, as will usually be the case, θ and ϕ are small fractions whose squares will be negligible, K' can be approximated by K'' where:

$$(1.03) \qquad K'' = 1 - 2\phi - 2\theta + \frac{\theta}{q} + \frac{\phi}{p} = 1 + \theta\left(\frac{1}{q} - 2\right) + \phi\left(\frac{1}{p} - 2\right)$$

ESTIMATION

An examination of K and K' tells the story concerning the effects of misclassification on estimates. *In the following discussion it will be assumed that p is less than $\frac{1}{2}$.* If p is greater than $\frac{1}{2}$ the roles of the two types of misclassification (i.e., of θ and ϕ) will be interchanged. *Moreover it will also be assumed that θ and ϕ are of the same order of magnitude*—this being the usual situation in practice. If θ is much larger than ϕ, the remarks may not apply.

If x/n is used as an estimate of p, a bias may be introduced by the misclassification. The bias is:

$$(1.04) \qquad B = q\phi - p\theta = (K - 1)p$$

so that K is a measure of the relative bias. Under the above assumptions, the chief bias will arise from misclassifications of the "non-problem" cases and the bias will get progressively worse as p becomes smaller and smaller. For example, if there is a 5% risk of misclassification in either direction ($\theta = \phi = .05$) and if there are 10% "problem" cases in the population ($p = .10$), then by (1.01):

$$K = 1 - .05 + \frac{.90}{.10}(.05) = 1.40$$

so that the proportion of problem cases will be overestimated by 40%. If estimates of θ and ϕ are available from the data used to construct the classification system, the estimates might be adjusted by dividing by K.

An alternative procedure would be to choose the critical point on the classification scale so that:

$$(1.05) \qquad \phi/\theta = p/q.$$

With this choice of the critical point unbiased estimates would be obtained. Unfortunately the value of p would not be known in advance so in practice (1.05) could only be approximated. Moreover, if (1.05) is used for the whole population the estimates for sub-populations might still be biased.

The misclassification not only introduces a bias but also affects the variance of the estimate. When an estimate is biased it is necessary to

distinguish clearly between the variance (which is the second moment about the expected value) and the mean square error (which is the second moment about the true value). With biased estimates the mean square error is a better index of efficiency.

The quantity K' measures the effect of misclassification on the variance of the estimate, however, it is easier to see the effects of misclassification if the approximate quantity K'' (1.03) is examined. It can be seen that as p becomes smaller the variance tends to increase. Note that the effect of misclassification of the "problem" cases (i.e. θ) is to *reduce* the variance. As p becomes smaller and smaller this effect becomes unimportant. On the other hand, the effect of misclassifications of the "non-problem" cases (i.e. ϕ) is to increase the variance and this effect increases as p becomes smaller and smaller. The accepted measure of efficiency is the reciprocal of $K' + D$ where

$$D = \frac{(q\phi - p\theta)^2 n}{pq}$$

It will be noted that the efficiency so defined will *decrease* as the sample size increases. Thus in a sample of 100 with $\theta = \phi = .05$ and $p = .10$, the value of K' is 1.34 and the value of D is 1.78. Therefore the index of efficiency, the reciprocal of $1.34 + 1.78 = 3.12$, will be 32%. A 32% efficiency has the interpretation that about $\frac{2}{3}$ of the original 100 observations are "lost" due to the misclassification.

The use of (1.05) to fix the critical point will not only help to unbias the estimates, but it will also greatly improve the efficiency. However, in many studies the critical point will necessarily be determined by other considerations than (1.05) and an appreciable bias may be introduced. As the foregoing examples indicate, relatively small misclassification rates can introduce very serious biases and lead to very low efficiencies. While in theory an adjustment can be made either by calculating K or by using (1.04) to calculate B, in practice the quantities θ and ϕ are required and only crude estimates of these quantities may be available. The research worker then has an unpleasant choice between an unadjusted estimate and a rather arbitrary adjustment. In this situation, it may be best to present both estimates together with a clear statement of the difficulty. The "standard" confidence intervals should *not* be used here.

SIGNIFICANCE TESTS

The principal interest of this paper lies in the comparison of samples from two populations (denoted by subscripts 1 and 2). To compare

the proportion of problem cases, the statistic

$$d = \frac{x_1}{n_1} - \frac{x_2}{n_2}$$

would be used.

An especially important practical situation is the one where the same classification system is used in both samples. It will be shown that for this situation, the significance test remains valid even though the misclassification is altogether ignored. The comfort that the research worker can draw from this result is somewhat mitigated by the fact that the significance test, though valid, becomes less powerful.

In the previous section it has been shown that the expected value of x when misclassification occurs is Kp and the variance is $nKp\ (1 - Kp)$ where $K = 1 - \theta - (q/p)\theta$. In terms of K' (see equation (1.02)) the variance of x is $npqK'$ so:

$$pqK' = Kp(1 - Kp)$$

If now, x_1 and x_2 are independent samples from two populations where the same classification system has been used, then

$$E(d) = \frac{n_1 p_1 K}{n_1} - \frac{n_2 p_2 K}{n_2} = K(p_1 - p_2)$$

and under the null hypothesis (i.e. $p_1 = p_2$) the expectation of d is 0 whether there is misclassification or not.

The variance of d under the null hypothesis will be:

$$V(d) = \frac{pqK'}{n_1} + \frac{pqK'}{n_2} = pqK'A$$

where

$$A = \frac{1}{n_1} + \frac{1}{n_2}.$$

If the misclassification is altogether ignored, the variance would be estimated by first estimating p by

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

and then estimating pq/N by $\hat{p}\hat{q}/(N - 1)$ where $N = n_1 + n_2$ so that

$$\hat{V}(d) = \hat{p}\hat{q}\frac{N}{N - 1}A.$$

While it is known that the above is an unbiased estimate of the variance where there is no misclassification, the question that now must be answered is: What is the effect of misclassification on this estimate?

When misclassification occurs

$$E\hat{p}\hat{q} = E\hat{p}(1 - \hat{p}) = E\hat{p} - E\hat{p}^2 = Kp - E\hat{p}^2$$

but

$$V\hat{p} = \frac{pqK'}{N} = E\hat{p}^2 - (E\hat{p})^2$$

so that

$$E\hat{p}^2 = (E\hat{p})^2 + \frac{pqK'}{N} = K^2p^2 + \frac{pqK'}{N}$$

Hence

$$E\hat{p}\hat{q} = Kp - K^2p^2 - \frac{pqK'}{N}$$

and

$$E\hat{V} = \frac{N}{N-1} A\left\{Kp(1 - Kp) - \frac{pqK'}{N}\right\}$$

while the actual variance is

$$V = pqK'A.$$

Since $Kp(1 - Kp) = pqK'$

$$E\hat{V} = \frac{N}{N-1} A\left\{pqK'\left(1 - \frac{1}{N}\right)\right\} = pqK'A.$$

so that even when misclasification is present, an unbiased estimate of the variance is obtained. This means that the existence of misclassification does not affect the validity of a significance test based on d, and since the ordinary Chi-square test is equivalent to a test based on d, it follows that the validity of the Chi-square test is also not affected by ignoring misclassification.*

We do not get off "scot-free" however. Although the tests are valid the power may be drastically reduced. In calculating the power of the test *effective* sample sizes of $1/K'$ times the actual sample sizes should be used. Table I presents values of the efficiency $(1/K')$ for representative values of θ, ϕ, and p.

*I am indebted to the referees for improvements in this proof.

TABLE I

THE INFLUENCE OF MISCLASSIFICATION ON EFFICIENCY

θ	p	ϕ .01	ϕ .02	ϕ .05	ϕ .10	ϕ .20
.01	.05	.856	.744	.542	.384	.261
	.10	.934	.871	.730	.587	.447
	.15	.963	.923	.825	.712	.587
	.20	.978	.951	.882	.797	.696
.05	.05	.884	.765	.552	.389	.262
	.10	.966	.898	.748	.597	.451
	.15	.995	.951	.846	.725	.593
	.20	1.008	.978	.903	.812	.703
.10	.05	.922	.793	.566	.395	.264
	.10	1.009	.934	.771	.610	.457
	.15	1.037	.989	.873	.743	.601
	.20	1.048	1.015	.932	.832	.713

A study of this table will give the research worker a better "feeling" for the effects of misclassification. It shows, for example, that when the "problem" case is a rare event then even small rates of ϕ (misclassification of "non-problem" cases) can result in the loss of about half of the data. Insofar as the significance test is concerned, the effect of misclassification of "problem" cases is not very great but it should be remembered that estimates will be much more seriously affected by this kind of misclassification.

If the test turns out to be significant we may wish to use d to estimate the difference in population proportions. This estimate will be biased by the factor K, however, and the previous remarks concerning estimation will apply here as well.

SMALL SAMPLES

Since we have been primarily interested in the situation where θ, ϕ, and p are small, it may occur to some readers that the normal distributions hitherto assumed might not be very good approximations in samples of a realistic size. Therefore it seems worthwhile to examine what happens in small samples in a simple special case.

Since the quantity v is going to be a fraction of u, it is relatively unimportant and for the sake of simplicity we will now consider only the quantities u and w, so:

$$x = u + w$$

As ϕ and p are small, it is appropriate to regard u and w as following Poisson distributions with expectations np and $(n - u)\phi$ respectively. As a further approximation we note that u is small relative to n, so that we shall take the expectation of w as simply $n\phi$. Under these assumptions the joint distribution of u_1 , u_2 , and w_1 , w_2 will be the product of four Poisson distributions. Thus the distribution of u_1 would be:

$$P(u_1) = \frac{e^{-n_1 p_1}(n_1 p_1)^{u_1}}{u_1!}$$

The joint distribution, $P(u_1 , u_2 , w_1 , w_2)$ can be rewritten in another form which will be more useful. Let $t = x_1 + x_2$. Then:

$$P(u_1 , u_2 , w_1 , w_2) = P(t)P(x_1 , x_2 \mid t)P(u_1 , w_1 \mid x_1)P(u_2 , w_2 \mid x_2)$$

where:

$$P(t) = \frac{e^{-A}A^t}{t!} \qquad A = n_1 p_1 + n_1\phi + n_2 p_2 + n_2\phi$$

$$P(x_1 , x_2 \mid t) = \frac{t!}{x_1!x_2!}P^{x_1}Q^{x_2} \qquad P = \frac{n_1(p_1 + \phi)}{n_1(p_1 + \phi) + n_2(p_2 + \phi)}$$

$$P(u_i , w_i \mid x_i) = \frac{x_i!}{u_i!w_i!}P_i^{u_i}Q_i^{w_i} \qquad P = \frac{n_i p_i}{n_i p_i + n_i\phi} \qquad i = 1, 2$$

The important point to note is that if there is no difference between populations the conditional distribution, $P(x_1 , x_2 \mid t)$ is a binomial distribution which does not depend on either p or ϕ, since:

$$P = \frac{n_1}{n_1 + n_2}$$

If there is no misclassification ($\phi = 0$), the same distribution of x_1 and x_2 is obtained. Consequently the validity of the significance test based on $P(x_1 , x_2 \mid t)$ is undisturbed by misclassification. This result agrees with the large sample result obtained previously.