Study of case control design and cohort design to determine the size of the samples according to the p-value, and confidence intervals

Abstract

In the introduction, we assume familiarity with the basic ideas and interpretation of hypothesis tests, including the p-value, and confidence intervals. These concepts, while appealing, are surprisingly subtle and can be disturbingly misleading. Even the language we have just used to describe the interpretation of a p-value is somewhat sloppy and imprecise. While we cannot do justice to the literature surrounding these topics, there are a few comments worth noting, having encountered the first hypothesis test in the book. The use of classical hypothesis testing, and associated p-values, has been roundly and deservingly criticized (Goodman and Royall, 1988). The p-value for a χ^2 test of independence does not represent the probability that the population Relative Risk is as far as or further from independence (RR=1) as the observed sample Relative Risk. The p-value is certainly not the probability of H_0, given the data—this is almost the error of equating $P(A|B)$ with $P(B|A)$.as the p-value depends rather on computing probabilities of possible observations, given H_0. Further, the p-value takes no account of the *power of the study* with regard to the hypothesis test, that is, the probability of accepting the null when it is actually false. Thus a small deviation from independence in a large study can have an identical p-value to a small study containing large deviations. In cynical moments, I find myself discarding a p-value as little more than an alternate measure of sample size since all null hypotheses will produce small p-values so long as enough data are collected. How should we then interpret the p-values we report here? We use them as informal measures of the compatibility of the data with the null hypothesis in question. This does not evade the criticisms outlined, or others for that matter (Goodman and Royall, 1988), but it does reinforce that they cannot be treated in a formal manner and certainly should not be subject to an arbitrary cutoff value such as 0.05. In addition, p-values arise from calculations based on a null hypothesis that is unlikely to be exactly true. As such, p-values can only be considered as approximations. This is further supported by the fact that they usually do not account for sources of error beyond sampling variation, nor for the impact of multiple comparisons (performing many tests on the same set of data, an action rarely acknowledged in single p-value calculations).One way to minimize the use of p-values is to focus on estimation of effects, rather than testing null values. We begin to look at this more closely in the next chapter. Uncertainty is often introduced into estimation through the use of confidence intervals. Although confidence intervals are subject to similar criticisms as p-values, they are better rough descriptors of the uncertainty involved in estimation because they avoid the more egregious misinterpretations associated with hypothesis testing, particular if more than one confidence level is used, as suggested in Chapter 3.3.Alternative inference procedures are available that should be given serious consideration. These include the use of likelihood intervals (see Chapter 13.1.1 for an introduction to the likelihood function) and Bayesian methods. A brief introduction to both of these techniques can be found in Clayton and Hills (1993). In light of the above comments on p-values, it is interesting to note that p-values tend to overstate the evidence against the null hypothesis, as compared to likelihood or Bayesian intervals, particularly when the p-value is greater than 0.01 (Berger and Sellke, 1987).

Introduction

1.1 Cohort designs

The logic used with population-based designs to investigate independence of D and E is not appropriate for data from a cohort design since it is not possible to estimate joint or marginal probabilities with this design. For example, consider the data in Table 5.3. Here the sample sizes in the random samples of unmarried and married mothers are preselected by the investigator; thus, the observed marginal frequency of unmarried mothers tells you nothing about the population frequency of this characteristic. Indirectly, the choice of the two sample sizes also determines the marginal frequency of low-birth weight infants so that the data again do not provide information regarding the occurrence of low birth weights in the population. By the same token, joint probabilities such as the probability of being an unmarried mother with a low-birth weight infant cannot be estimated from cohort data.

Nevertheless, it is still possible to investigate independence of D and E with cohort data through the equivalent formulation of independence as $P(D|E)=P(D|\text{not } E)$. Both of these conditional probabilities are immediately estimable from the two distinct exposure category samples, and the issue of independence then simplifies to the comparison of two separate population proportions or probabilities.

Specifically, we write the null hypothesis

$$H_0 : D \text{ and } E \text{ are independent} \Leftrightarrow P(D|E) = P(D|\text{not } E)$$

Page 63

For simplicity write $p_1=P(D|E)$ and $p_2=P(D|\text{not } E)$. Using the notation of Table 6.1, we estimate p_1 and p_2 by

$$\hat{p}_1 = \frac{a}{a+b} = \frac{a}{n_1}$$
$$\hat{p}_2 = \frac{c}{c+d} = \frac{c}{n_2}$$

where n_1 and n_2 are the sample sizes of the E and $\bar{E}$ samples, respectively. The two estimates $\hat{p}_1$ and $\hat{p}_2$ are random variables whose sampling distributions can be approximated by Normal distributions when both n_1 and n_2 are large. The expectation of the approximating Normal sampling distribution for $\hat{p}_1$ is p_1, with variance given by $p_1(1-p_1)/n_1$, and an analogous result holds for the approximate Normal sampling distribution of $\hat{p}_2$. Thus, the difference between the two estimates, $\hat{p}_1 - \hat{p}_2$, has an approximate Normal sampling distribution with expectation p_1-p_2 and variance $[p_1(1-p_1)/n_1]+[p_2(1-p_2)/n_2]$.

Now, if we assume the null hypothesis, that D and E are independent, is true, then $P_1=P_2$. We can then use the difference of our estimates, $\hat{p}_1 - \hat{p}_2$, as a measure of the dependence of D and E. If D and E are independent—that is, H_0 is true—then the parameters of the approximate Normal sampling distribution of $\hat{p}_1 - \hat{p}_2$ simplify; its expectation is then 0 and its variance is given by $p(1-p)[(1/n_1)+(1/n_2)]$, where $p=p_1=p_2$. The latter variance can be estimated by $\hat{V} = \hat{p}(1-\hat{p})[(1/n_1)+(1/n_2)]$, with $\hat{p} = (a+c)/n$, the observed proportion of Ds in the whole sample.

A reasonable test statistic is thus given by $(\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{V}}$, which is expected to follow a standard Normal sampling distribution if D and E are independent. If this statistic is large, there is evidence that D and E are associated. As before, "large" is interpreted in terms of the sampling distribution assuming independence. Alternatively, we can compute the square of this statistic, given by $(\hat{p}_1 - \hat{p}_2)^2/\hat{p}(1-\hat{p})[(1/n_1)+(1/n_2)]$, which is approximately $\chi^2_{(1)}$ under H_0.

Some simple algebra shows that this test statistic is, in fact, equivalent to the χ^2 test statistic described in Section 6.1 for population-based data.

Table 6.3 *Possible data from a cohort study* $(n_E = n_{\bar{E}} = 100)$ *of a mother's marital status and low birthweight*

		Birthweight		
		Low	Normal	
Marital status at birth	Unmarried	12	88	100
	Married	5	95	100
		17	183	200

Thus,

$$\frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = \frac{(ad-bc)^2}{n_1^2 n_2^2} \times \frac{n n_1 n_2}{(a+c)(b+d)}$$
$$= \frac{n(ad-bc)^2}{n_1 n_2 (a+c)(b+d)}$$
$$= \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)},$$

as claimed.

To illustrate the computation in this setting, we use the data of Table 5.3, repeated in Table 6.3 for convenience, which yields the observed test statistic of $200(12\times95-88\times5)^2/100\times100\times17\times183-98{,}000{,}000/31{,}110{,}000-3.15$, with an associated p-value of 0.08.

1.2 Case-control designs

As for cohort designs and for the same reasons, it is not possible to estimate either joint or marginal probabilities from a case-control design. Further, with case-control data, we cannot obtain information about the conditional disease proportions, *P(D|E)* and *P*(*D*|not *E*), either, since the disease frequencies in the data are determined by the sample sizes of cases and controls as pre-specified by the investigator. In fact, the only probabilities that are estimable are exposure probabilities, conditional on disease status; that is, $P(E|D)$, $P(E|\bar{D})$. Here H_0, that D and E are independent, can then be given as $P(E|D) = P(E|\bar{D})$, an alternate specification of independence. Thus, in an analogous fashion to that used for cohort designs, we can assess the hypothesis of independence by comparing the observed sample frequency of exposed individuals among cases against that among controls. Using identical algebra to that used in Section 6.2, this comparison the yields the exact same χ^2 statistic used to test independence in both population-based and cohort designs. In sum, although based on differing justifications, we see that the identical χ^2 test of independence applies to each of the three designs we have considered.

Using the data given in Table 5.4, repeated in Table 6.4 for convenience, we illustrate the by-now familiar calculation of the χ^2 statistic for case-control data. Here, the observed test statistic is given by $200(50\times72-50\times28)^2/100\times100\times78\times122 = 968{,}000{,}000/95{,}160{,}000=10.17$. This is a very large value in terms of the $\chi^2_{(1)}$ distribution and yields a p-value of 0.002. Thus, the case-control data provides quite striking evidence that a mother's marital status is related to the possibility of having a low-birthweight infant. In the next section, we explain the marked difference between the results of the χ^2 test for data generated by each of the three design strategies applied to the same population and with the same sample size.

1.2.1 Comparison of the study designs

Table 6.5 summarizes the results of the χ^2 test of independence on the data from Tables 6.2 to 6.4. Note that the results of the three tests are not entirely incompatible with each other. However, while both the population-based and cohort data are merely suggestive of an association between a mother's marital status and infant birthweight, the case-control data appear to provide sufficient evidence to reject the notion that these two factors are unrelated.

The explanation of the differences between the designs can be found in terms of the *power* of the χ^2 test, that is, the probability that the test will reject independence given a population association between the two factors of interest. Although we cast doubt on the value of hypothesis testing in Section 6.1.1, the power of the χ^2 test remains a useful proxy for the amount of information in a data set regarding the question of independence of D and E. As with most hypothesis tests, the power of the χ^2 test depends on the extent of the true unknown population association and the sample size. This does not explain the different results generated by the three study designs since these factors are the same in each case.

Additional factors influence the power of the χ^2 test, namely, the *balance* of the two marginal totals for D and E, respectively. First, let us compare the population-based and cohort designs. Recall that, for either design, the χ^2 test is based on the (square of the) statistic $(\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{V}}$ where $p_1=P(D|E)$, $p_2 = P(D|\bar{E})$, and $\hat{V} = \hat{p}(1-\hat{p})[(1/n_1)+(1/n_2)]$ where n_1 and n_2 are the sample sizes for the Es and Es, respectively. If the null hypothesis is false, then D and E are related, and $p_1-p_2\neq0$. The power of the χ^2 test increases with the size of p_1-p_2, but, in a given population, this difference is fixed and therefore not influenced by choosing either a population-based or cohort design. Similarly, near the null hypothesis, $p=p_1=p_2$ is also fixed. The variance term is, however, affected by the design through the term $[(1/n_1)+(1/n_2)]=n/n_1n_2$. As this term decreases, the precision of our estimate $\hat{p}_1 - \hat{p}_2$ increases, and the χ^2 test statistic also increases. With the total sample size n fixed, n/n_1n_2 is minimized when n_1n_2 is maximized, which occurs when $n_1=n_2=n/2$. From this we can infer that, for a cohort design with fixed sample size n, the best sample size allocation, in terms of statistical power of the χ^2 test, is to take $n_1=n_2=n/2$. Further, since the sample sizes of Es and Es are random in a population-based design, they will be essentially determined by the population $P(E)$, and will almost always be unequal, even if $P(E)=0.5$. Hence, for large samples, a population-based design always yields a less powerful χ^2 test than a cohort design with equal sample sizes of Es and Es.

By the same token, comparison of the case-control design to the population-based design can be considered in terms of p_1-p_2, where now $p_1=P(E|D)$ and $p_2 = P(E|\bar{D})$. According to the logic of the last paragraph, the most powerful choice of sample sizes for the case-control design is $n_1=n_2=n/2$, where n_1 and n_2 are now the sample sizes for the Ds and $\bar{D}$s, respectively. And, if equal (large) sample sizes of cases and controls are used, the case-control design always leads to a more powerful χ^2 test than a population-based design of the same population.

Finally, the cohort and case-control designs are compared, assuming that both use the optimal equal allocation of the overall sample size to their respective two random samples. This removes the influence of sample size so that differences in power now depend solely on the expected value of the part of the χ^2 statistic given by $(\hat{p}_1 - \hat{p}_2)/\sqrt{\hat{p}(1-\hat{p})}$, where $\hat{p} = (\hat{p}_1 + \hat{p}_2)/2$ since $n_1=n_2$. In large samples, this expectation is approximately $d = (p_1 - p_2)/\sqrt{p(1-p)}$, where for cohort designs $p_1=P(D|E)$, etc., and for case-control designs $p_1=P(E|D)$, etc., and in either design $p=(p_1+p_2)/2$. The power of the χ^2 test will be greater when d is larger, and this scaled difference grows as the average of p_1 and p_2, that is, p, gets closer to 0.5. Figure 6.1 illustrates this graphically for three different population Odds Ratios. Thus, in comparing a cohort to a case-control design, the greater power will belong to the design for which the average, p, of the relevant conditional probabilities lies closer to 0.5. That is, if $[P(E|D) + P(E|\bar{D})]/2$ is nearer to 0.5 than $[P(D|E) + P(D|\bar{E})]/2$, then the case-control design will have higher power than the cohort design for any population Odds Ratio that differs from 1, and vice versa. With $n_1=n_2$, $P(E)$ closer to 0.5 than $P(D)$ implies that $[P(E|D) + P(E|\bar{D})]/2$ is nearer to 0.5 than $[P(D|E) + P(D|\bar{E})]/2$, and that therefore the case-control design is more powerful with large samples; when $P(E)$ is closer to 0.5 than $P(D)$, the converse of this statement is true by the same reasoning.

In summary, we have learned that for large samples,

• In a cohort design with a fixed sample size, the χ^2 test of independence is most powerful when the exposed and unexposed samples are of equal size.

• A cohort design with equal samples of exposed and unexposed yields a more powerful χ^2 test of independence than a population-based study with the same overall sample size.

• In a case-control design with a fixed sample size, the χ^2 test of independence is most powerful when the case and control samples are of equal size.

• A case-control design with equal samples of cases and controls yields a more powerful χ^2 test of independence than a population-based study with the same overall sample size.

• When *P(E)* is closer to 0.5 than *P(D),* the case-control design with equal samples of cases and controls will give a more powerful χ^2 test of independence than the cohort design with equal numbers of exposed and unexposed.

• When *P(D)* is closer to 0.5 than *P(E)* then the cohort design with equal numbers of exposed and unexposed will give a more powerful χ^2 test of independence than the case-control design with equal samples of cases and controls.

These conditions all point to greater power being achieved when *both* disease and exposure marginal frequencies are closer to being balanced. In either a cohort or case-control study, one marginal can be exactly balanced by design, and the relative size of *P(D)* and *P(E)* in the population determines whether greater balance can be gained in the other marginal by one design or the other. Let us now look back at Tables 6.2 to 6.4 to observe how marginal balance differences explain the comparison of the χ^2 test statistics of Table 6.5. The exposure (marital status) marginal is exactly balanced in Table 6.3, and the outcome (birthweight) marginal is slightly better than in Table 6.2 from a population-based design. The power gain from the cohort design here arises solely from the sample allocation term of the χ^2 test statistic, namely n/n_1n_2, which is 0.02 for any cohort design with $n_1=n_2=100$, and 0.024 (=200/59×141) for the specific population-based outcome of Table 6.2, indicating that we can expect the slight increase in power reflected in Table 6.5. With completely balanced cohort and case-control designs, the differences in power are entirely due to variation in *d,* as discussed above. For the cohort design, d=0.240 (with $p_1=P(D|E)$=0.12 and $p_2 = P(D|\bar{E}) = 0.05$), whereas, for the case-control design, d=0.352 (with $P_1=P(E|D)$=0.5 and $p_2 = P(E|\bar{D}) = 0.28$). This confirms that we can anticipate a substantial increase in power by using a case-control design here, as we see in Table 6.5.

Note that these power comparisons can change if you choose a case-control design or cohort design with unequal sample allocations. For example, even when *P(D)* is much smaller than *P(E),* it is possible that the case-control design will yield less power than the cohort design if you allocate the total sample poorly. Further, remember that increasing the total sample size will increase power for all designs (again assuming a sensible allocation of this sample size in both the case-control designs and cohort designs). Thus, even if the available sample size of cases is limited in a case-control design (say, to 100, for example), it still adds power if we sample more than 100 controls as compared to balancing the sample sizes at 100. The gain in power comes from the decline in the sample size factor, $ssf=n/n_1n_2$, as n increases, even though n_1, the number of cases, stays fixed. If $n_2=kn_1$, say, then $ssf=(k+1)/kn_1$. The relative size of ssf for k=1 compared to k>1 is $R=2k/(k+1)$, with the value of R then reflecting the ratio of the sample size factor with k controls per case to that with one control per case. Figure 6.2 plots R against k, the ratio of the number of controls to number of cases. With a fixed number of cases, the figure shows the growth in the sample size factor—and thus the power of the χ^2 test—as you increase the number of controls selected per case; however, as a rule of thumb, it is clear that you gain relatively little by adding extra controls, after you have four times as many controls as cases. The primary gain in power comes from increasing the number of controls per case from 1 to 4.

2.3 Comments and further reading

The power comparisons between the various designs, with resulting sample size implications given in Section 6.3.1, are appropriate at the null, that is, assuming independence between *D* and *E*. Different recommendations for sample size allocation are necessary when estimating a relationship away from the null. We return to this briefly in Chapter 7.1.1. Also, note that the χ^2 test is not immediately applicable

Figure 6.2 *Relative size of sample size factor, R, compared to the ratio of number of controls to cases, k, for a case-control design.*

to case-control data with risk-set sampling except under restrictive conditions. For a case-cohort design, the χ^2 test of independence can be directly applied, not to the data as represented in Table 5.7, but to the version where the cohort sample is modified by removal of any cases.

2.3.1 Alternative formulations of the χ^2test statistic

For population-based designs, the *a* entry in Table 6.1 is a random variable with expectation $E(a)=nP(D\&E)$. Be careful not to confuse “Expectation” and “Exposure.” If independence of *D* and *E* is assumed, then $E(a)=nP(D)P(E)$. Similar formulae can be developed for the *b, c,* and *d* entries. It is then easy to show that the χ^2 statistic can also be written as

$$\chi^2 \text{ statistic } = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \qquad (6.1)$$

where *i* and *j* index the four cells of the 2×2 tables so that $O_{11}=a$ etc., and E_{ij} denotes the estimated expectation of the relevant cell *under the assumption of independence*. That is, E_{11} is the estimate of $E(a)=nP(D)P(E)$ given by $E_{11}= n[(a+b/n)(a+c/n)]$. Equation 6.1 also holds under both cohort and case-control sampling schemes.

Yet another derivation of the χ^2 test for independence can be based on the following argument. Very little, if any information, about the relationship between *D* and *E* can be gleaned from the marginal entries of Table 6.1, that is, the row and column totals. It is the “interior” entries of the table that tell us about the strength of association of *D* and *E*. We may as well then assume that the marginal totals are *fixed* at their observed values, and then try to determine whether the *a, b, c,* and *d* entries suggest possible independence or otherwise. With fixed marginal totals, analysis of the data from any of the three designs is then identical (although, of course, the different designs will generate different marginals which have power implications as we have seen). Further, if the marginals are fixed, only one piece of information remains random within Table 6.1; that is, if we treat *a* as a random variable, the other entries, *b, c,* and *d,* are all determined once *a* is known. Thus, questions about the relationship between *D* and *E* can all be couched in terms of the properties of the random variable *a* once we assume that the marginal's are fixed and known. In fact, the random variable *a* then follows what is known as the *non-central hypergeometric distribution,* parameters of which are determined by the known marginal totals and the unknown population Odds Ratio. In the special case of independence of *D* and *E,* this distribution simplifies and is known simply as the *hypergeometric distribution*. With independence assumed, the expectation and variance of a can be simply described in terms of the marginal totals; specifically, $E(a)=(a+b)(a+c)/n$, and $Var(a)=(a+b)(c+d)(a+c)(b+d)/n^2(n-1)$. When n is large, the hypergeometric distribution is well approximated by a Normal distribution with the same expectation and variance. Thus under the null hypothesis, , of independence, the random variable $[a - E(a)]/\sqrt{Var(a)}$ should approximately follow a standard Normal distribution, or equivalently, $(a-E(a))^2/Var(a)$ is approximately $\chi^2_{(1)}$. This then provides the basis for a test of independence of D and E. In fact, the test statistic, $(a-E(a))^2/Var(a)$, differs from the χ^2 statistic, $n(ad-bc)^2/(a+b)(c+d)(a+c)(b+d)$, only in that the term n in the numerator is replaced by $(n-1)$. This is irrelevant when n is large, and the two approaches will then give almost identical results. However, this difference between n and $(n-1)$ in these two versions of the χ^2 statistic will have important implications when we study the combination of many 2×2 tables with small sample sizes in Chapters 9 and 16.

2.3.2 When is the sample size too small to do a χ^2 test?

In discussing the χ^2 test of independence of D and E, we frequently referred to "large" samples or "large" n in order to invoke the approximation of the sampling distribution of the test statistic by the $\chi^2_{(1)}$ distribution. Just how large does n have to be? In fact, the quality of the approximation does not depend solely on n; extensive examination has determined that it is accurate so long as the expectation (assuming independence) of each entry inside Table 6.1 is greater than 1 (Larntz, 1978). We can check this by using estimates of these expectations for each entry in the 2×2 table; that is, by examining whether $(a+b)(a+c)/n$ is greater than 1 for the a entry, and so on. For example, Tables 6.2 to 6.4 meet these criteria easily.

The exact sampling distribution of cell entries can be used to construct a test of independence when the sample size is so small that use of the approximating $\chi^2_{(1)}$ distribution is questionable. With the assumption of fixed marginals, the relevant exact distribution is the hypergeometric as noted in Section 6.4.1, whose use as the null sampling distribution of a leads to the *Fisher exact test.* Further discussion of this test can be found in either Fleiss (1981) or Breslow and Day (1980). For either cohort or case-control designs, an alternative exact test is based on the binomial distributions for each of the two samples generated by the design. This exact test has somewhat greater power than Fisher's exact test (D'Agostino et al., 1988). The widespread availability of such exact tests precludes the need to use a continuity correction to improve the adequacy of the χ^2 approximation; if we face a 2×2 table where use of the continuity correction makes a noticeable difference, then proceed with an appropriate exact test and avoid use of the χ^2 test altogether.

For case-cohort data, the sample Odds Ratio can be used to estimate the Relative Risk although the variance estimate in Section 7.1.2 is generally incorrect because of possible overlap between cases and controls—that is, the possibility that the control group contains individuals who are subsequently sampled as cases. While correcting the variance estimate is straightforward, Sato (1992a) describes a modified estimator of the Odds Ratio that takes into account the information in the overlapping group to improve precision (a slightly different estimator is relevant if one knows the size of, and the number if cases in, the larger cohort from which the case-cohort data was sampled (Prentice, 1986; Sato, 1992a). If there is very little overlap, these modifications can be ignored and the methods of this chapter directly applied, although this essentially returns us to the rare disease setting. In addition, if the "overlapping" cases are removed from the cohort sample, the methods of Section 7.1 directly apply to estimation of the Odds Ratio as for traditional case-control studies.

Conclusion

In this study, we have considered quantifying uncertainty in a study of a particular design and sample size(s). Turning these techniques on their heads, we can determine the size of the sample(s) required to achieve a given level of precision for a specific design. Such calculations are referred to as *sample size planning*. There is a substantial literature on this topic, and we refer to Schlesselman (1982) and Greenland (1988) as good places to start. There are also chapters on this topic in many other books including Woodward (1999) and Newman (2001). For a straightforward introduction and review, see Liu (2000). I find sample size computations somewhat artificial. Usually, available resources for a study are constrained, and sample size calculations are often used to justify the value of a study, given fixed resources, as compared with precision assessment driving appropriate fund allocations. Further, sample size planning rarely accounts for all sources of error, some of which may be a far greater threat than sampling variability. For example, it may be more effective to expend a greater fraction of resources ensuring the quality of measurement of exposure and disease than to merely increase the sample size for a study with inaccurate data. It is particularly dangerous to blindly resort to sample size tables without fully understanding the statistical nuances of a planned design and analysis strategy.

As indicated at the beginning of Chapter 2, we have been assuming exact measurement of both disease and exposure so far, particularly in this chapter. However, bias introduced by both systematic and random errors in measurement of these quantities leads often to, far greater distortion of a disease-exposure association than is introduced by sampling. Thus, any epidemiological study must assess the possible impact of measurement error before drawing definitive conclusions. Even relatively small amounts of measurement error can have major effects on estimation of a measure of association. The direction of the association can be reversed, as is easily seen if we imagine.

Reference

R. Caruana and V. de Sa. Benefitting from the variables that variable selection discards. *JMLR*, 3: 1245–1264 (this issue), 2003.

I. Dhillon, S. Mallela, and R. Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *JMLR*, 3:1265–1287 (this issue), 2003.

T. G. Dietterich. Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, USA, 2nd edition, 2001.

T. R. Golub et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

G. Forman. An extensive empirical study of feature selection metrics for text classification. *JMLR*, 3:1289–1306 (this issue), 2003.

T. Furey, N. Cristianini, Duffy, Bednarski N., Schummer D., M., and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16:906–914, 2000.

A. Globerson and N. Tishby. Sufficient dimensionality reduction. *JMLR*, 3:1307–1331 (this issue), 2003.

Y. Grandvalet and S. Canu. Adaptive scaling for feature selection in SVMs. In *NIPS 15*, 2002.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer series in statistics. Springer, New York, 2001.

T. Jebara and T. Jaakkola. Feature selection and dualities in maximum entropy discrimination. In *16th Annual Conference on Uncertainty in Artificial Intelligence*, 2000.

K. Kira and L. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *International Conference on Machine Learning*, pages 368–377, Aberdeen, July 1992. Morgan Kaufmann.

R. Kohavi and G. John. Wrappers for feature selection. *Artificial Intelligence*, 97(1-2):273–324, December 1997.

D. Koller and M. Sahami. Toward optimal feature selection. In *13th International Conference on Machine Learning*, pages 284–292, July 1996.

Y. LeCun, J. Denker, S. Solla, R. E. Howard, and L. D. Jackel. Optimal brain damage. In D. S. Touretzky, editor, *Advances in Neural Information Processing Systems II*, San Mateo, CA, 1990. Morgan Kaufmann.