Random sampling and study population *with regard* the relationship between birth weight and infant mortality

Abstract

We almost always have to deal with samples with limited and varying information. With a random sample of births, how do we examine the relationship between birth weight and infant mortality? What kind of accuracy can we expect and how does it depend on the size of the sample? Does it matter how we select the births to be sampled? We must confront the following issues:The methods for addressing these questions depend on how we draw the random sample from the population. In this chapter, we consider simple forms of random sampling and their broad impact on the answers to these questions. In the following chapters, we discuss both the statistical significance of an observed sample association and estimation of various measures of association.

Introduction

Before discussing study designs, we describe nested components of the population of interest. The *Target Population* refers to the population to which we would like to apply our estimates and inferences regarding the relationship between disease and exposure. Sometimes, it can be extremely difficult to sample directly from the Target Population; in such cases, there is often a convenient subgroup of the population for which appropriate sampling frames are available. We call this subgroup the *Study Population,* the population from which we are able to sample. Finally, the *Sample* comprises the actual sampled individuals from the Study Population for whom we collect data on disease, exposure, and other factors. Figure 5.1 is a schematic of these three groups. Note that the figure is not intended to be representative of scale. Typically the Study Population is a very large fraction of the Target Population, whereas the Sample is extremely small relative to the Study and Target Populations. For example, in many studies, a telephone interview may be used to collect information on study subjects. In these surveys, the Study Population comprises individuals in families that possess a residential telephone. As another example (Section 3.5), the Target Population might be the general community from which one might be tempted—with great risk as we have seen—to use individuals with hospital or clinic records as a convenient Study Population. The exercises at the end of this chapter contain several examples of epidemiological studies of various designs, illustrating possible choices of Study Populations within a given Target Population.

Intelligent choice of Study Populations will help us investigate our primary interest, the relationship between E and D. While selecting an appropriate Study Population is often predicated on the availability of sampling frames and other sampling mechanics, there are situations where the choice is based on the need to obtain valid comparisons for estimating an effect measure. This is particularly true in the cohort and case-control designs of Sections 5.2 and 5.3. On the other hand, differences between the Target Population and Study Population introduce *selection bias* in our results if the Study Population is not representative of the Target Population *with regard to the disease-exposure relationship* of concern. This does not necessarily require that the Study Population is representative of all aspects of the Target Population. However, when random sampling is used, differences between the Sample and

the Study Population are entirely due to random or sampling variation associated with the sampling technique employed. We can then use statistical methods to assess and describe these differences based on a detailed understanding of sampling procedures and variation. If the study sample is not selected randomly, we can treat the data in the same manner but without the same confidence in the calculations. Substantial bias can be introduced at this point if factors, often unmeasured or unknown, influencing the sample selection is associated with exposure and disease.

How do we usually obtain a random study sample from the Study Population? Three basic forms of sampling schemes are most commonly used in epidemiological studies. In each, we restrict attention to the association between the presence and absence of two binary factors, the outcome *D* and the exposure *E,* since the basic concept and the primary statistical impact of the designs are all captured even in this simplest scenario. Note that the sample data from any of the designs can be summarized in the form of a 2×2 contingency table as illustrated in Table 5.1. The rest of this chapter describes the three typical designs in terms of their statistical characteristics, determined by how study participants are sampled.

1 Population-based studies

The main steps of a *population-based* design are simply:

1. Take a simple random sample of size *n* from the Study Population.

2. Subsequently, measure the presence and absence of both *D* and *E* for all sampled individuals.

Note that the word "subsequently" here refers to the order of sampling individuals and measuring the factors, *D* and *E,* for the sample; there are no requirements on the chronological timing of events that determine *D* and *E* relative to the time of sampling. A further sub-classification of the design is often used to differentiate the timing of measurements on *D* and *E*. Specifically, Rothman and Greenland (1998) refer to a *prospective* study as one in which measurement of exposure is made on an individual prior to the occurrence and thus measurement of disease. Conversely, in a *retrospective* study, measurement of exposure occurs after an individual's disease status has been determined. A population-based study is often loosely called a cross-sectional study, but I prefer the former name as the latter suggests that measurement of *D* and *E* always coincides with sampling.

Whether a study is prospective or retrospective is not relevant to the study design and therefore not of immediate concern to the development of statistical properties. However, this classification may have considerable influence on the quality and validity of exposure measurement. For example, exposure assessment in a retrospective design must (1) evaluate the relevant risk levels in place *before* disease, and not after, and (2) ensure that measurements are not influenced by an individual's disease status. Note that prospective measurement of *D* may require a 10- or 20-year follow-up period after sampling.

The various types of population probabilities that may be of interest to the investigator can be classified as follows:

• Joint probabilities: $P(D \,\&\, E)$, $P(D \,\&\, \bar{E})$, $P(\bar{D} \,\&\, E)$, $P(\bar{D} \,\&\, \bar{E})$

• *Marginal* probabilities: $P(D), P(E), P(\bar{D}), P(\bar{E})$

• *Conditional* probabilities: $P(D|E), P(D|\bar{E}), P(E|D), P(E|\bar{D})$.

Each of these kinds of probabilities can be estimated using data generated from a population-based sample: estimates are given by the observed proportion of the simple random sample that corresponds to the population probability of interest.

In Chapter 3, we introduced data on the role of a mother's marital status or her baby's birthweight on subsequent infant mortality. A natural follow-up question is the extent to which the impact of marital status on infant mortality might be explained by birthweight. That is, it is plausible that unmarried women may receive poorer nutrition and prenatal care than married mothers-to-be, and thus deliver lower birth weight babies on average, which, in turn, would raise the risk of infant mortality substantially. To examine the relationship between marital status and birth weight, an investigator needs to collect data on these two factors in the population of interest.

Suppose that a sample size of 200 has been chosen for a population-based study. That is, a simple random sample of 200 births is selected from the Study Population. Table 5.2 shows a possible outcome of such a study. From this population-based data, we can then estimate:

• Joint probabilities, such as $\hat{P}$(unmarried mother and low birth weight infant)= 7/200=0.035

• Marginal probabilities, such as $\hat{P}$(low birth weight infant) =14/200=0.07

• Conditional probabilities, such as $\hat{P}$(low birth weight infant/unmarried mother)= 7/59=0.119, or $\hat{P}$(low birth weight infant/married mother)=7/141 = 0.050.

Sensible estimates can be obtained of the Relative Risk, Odds Ratio, Excess Risk, and Attributable Risk for a low-birth weight infant associated with the mother's marital status using the relevant estimates of the conditional probabilities $P(D|E), P(D|\bar{E})$, etc. in the definitions of a particular effect measure. We will discuss these estimates in more detail in the next chapter. For now, we see that

• $\widehat{RR}$=(7/59)/(7/141)=2.39

• $\widehat{OR}$=[(7/59)/(52/59)]/[(7/141)/(134/141)]=2.58

• $\widehat{ER}$=(7/59)–(7/141)=0.069

• $\widehat{AR}$=[(14/200)–(7/141)]/(14/200)=0.29.

Note the (by-now familiar) slightly higher value for the Odds Ratio as compared with the Relative Risk. The estimate of the Attributable Risk suggests that close to 30% of low birth weights in the population are attributable to the mother's marital status. Maternal marital status is presumably not casually associated with low birth weight but a proxy for poorer prenatal care and nutrition, as suggested earlier.

5.2

The primary feature of a *cohort* study is that sampling is carried out separately for subpopulations at different exposure levels, leading to distinct cohorts. The main steps of a cohort design are:

1. Identify two subgroups of the population on the basis of the presence or absence of *E*.

2. Take a simple random sample from each of these two subgroups (that is, the *E*s and not *E*s) *separately,* of sizes n_E and $n_{\bar{E}}$, respectively.

3. Measure subsequently the presence and absence of *D* for individuals in both random samples.

As for population-based studies, the timing and manner of measurement of the two factors *D* and *E* are not pertinent to the sampling characteristics of a cohort design. The key statistical property of the design is the separate identification and sampling of the exposure groups. When and how *D* and *E* are measured are important considerations in assessing the potential accuracy and bias in disease and exposure measurement, but are not germane to the direct statistical impact of the design itself.

Note that the investigator has to pre-specify the sample sizes for the two separate samples taken from the exposure groups. This division of the overall sample size is important in determining the amount of information that a cohort study yields on the disease-exposure relationship, as we shall discuss further. For an extreme example, if one exposure group is allocated a very small sample size, then there will be little information available on the disease-exposure relationship.

Table 5.3 shows a possible outcome of a cohort study using the same population as for the population-based design in Section 5.1.1. Here, we have selected two random samples, each of size 100, the first from the population of unmarried mothers and the second from married mothers. This design assumes that, prior to sampling, one is able to divide the population by marital status into two distinct sampling frames.

Data arising from such a cohort design have the following implications for estimation of population probabilities:

• Joint probabilities cannot be estimated—clearly, frequencies of joint characteristics such as unmarried mothers with low birthweight babies are artificially influenced by the exact allocation of the number of unmarried mothers sampled from the total sample of 200.

• Marginal probabilities are not estimable for the same reason.

• Only conditional probabilities that condition on exposure status can be estimated, such as $\hat{P}(D|E) = \hat{P}$(low birth weight infant|unmarried mother)=12/100 = 0.120, or $\hat{P}(D|\bar{E}) = \hat{P}$ (low birth weight infant|married mother)=5/100= 0.050.

The estimable conditional probability estimates provide essentially the same picture as those yielded by the population-based study of the same population (although the precision of these estimates may not be the same, but this is getting ahead of us).

Although only some basic conditional probabilities are estimable from a cohort design, these are fortunately the basic building blocks of the Relative Risk, Odds Ratio, and the Excess Risk. The Attributable Risk is not directly estimable from a cohort study because we cannot estimate *P(E)*. From Table 5.3 we can estimate

- $\widehat{RR}$=(12/100)/(5/100)=2.40
- $\widehat{OR}$=[(12/100)/(88/100)]/[(5/100)/(95/100)]=2.59
- $\widehat{ER}$=(12/100)–(5/100)=0.070.

Again, these estimates are compatible with those provided by the population-based data from the same population.

5.3 Disease-based sampling—case-control studies

A case-control study has the same specifications as a cohort study, except that the roles of *E* and *D* are reversed. Separate samples are thus selected from cases *(D)* and nondiseased individuals or controls $(\bar{D})$. The main steps of the design are:

1. Identify two subgroups of the population on the basis of the presence or absence of *D*.

2. Take a simple random sample from each of these two subgroups (that is, the *D*s and not *D*s) *separately,* of sizes n_D and $n_{\bar{D}}$, respectively.

3. Measure subsequently the presence and absence of *E* for individuals in both random samples.

As for cohort designs, the investigator must prespecify the number of cases and controls selected in the two separate random samples. Table 5.4 describes a possible outcome of a case-control study of mother's marital status and infant birthweight using samples of 100 cases *(D)* and 100 controls (not *D*). Here, implementing the design involves sampling first 100 low birthweight infants and then taking a further random sample of 100 normal birthweight infants. This again assumes that two sampling frames, one of low birth weight infants in the population and the other of normal birth weight infants, are accessible to the investigator.

For similar reasons as in cohort designs, only a limited set of probabilities can be estimated using case-control data:

• Joint probabilities cannot be estimated—frequencies of joint characteristics are again artificially influenced by the exact allocation of the number of low birthweight babies sampled from the total sample of 200.

• Marginal probabilities are not available for the same reason.

• Only conditional probabilities that condition on outcome status, here infant birth-weight, can be estimated such as $\hat{P}(E|D) = \hat{P}$(unmarried mother|low birth-weight infant)=50/100=0.500, or $\hat{P}(E|\bar{D}) = \hat{P}$(unmarried mother|normal birthweight infant)=28/100=0.280.

At first glance, it seems unlikely that we can estimate any measure of association from a case-control design. This is indeed partly true in that it is impossible to estimate the Relative Risk or the Excess Risk with case-control data. However, we can directly estimate the Odds Ratio for *E* associated with *D,* given by *[P(E|D)]/ [P(*not *E|D)]/[P(E|*not *D)]/[P(*not *E|*not *D)],* and then take advantage of the fact that this is identical to the Odds Ratio for *D* associated with *E* (using the symmetry of the roles of disease and exposure in the definition of the Odds Ratio that we highlighted in Section 4.4). Thus, from Table 5.4,

$$\widehat{OR} = [(50/100)/(50/100)]/[(28/100)/(72/100)] = 2.57,$$

compatible with the estimates provided by the population-based and cohort data.

In a situation where the outcome *D* is rare in both exposed and unexposed populations, the Odds Ratio will closely approximate the Relative Risk so that the case control estimate of the Odds Ratio can be used as an approximate estimate of the Relative Risk. It was, in part, this observation—that case-control studies can still be used to estimate Relative Risks in rare disease settings (Cornfield, 1951)—that led to their increased popularity as a study design over the past 50 years. The first modern use of the design was a study of the effect of reproductive history on the incidence of breast cancer (Lane-Claypon, 1926). The next section shows that the rare disease assumption is unnecessary for estimating the Relative Risk or Relative Hazard from case-control data if clever adjustments are made in the sampling of controls.

The Attributable Risk also appears to be inestimable from a case-control design. However, in the rare disease setting, we can again obtain an approximation; to see this, we first need some algebraic work to derive an alternative formulation for *AR*. Recall from section 4.7 that $AR = [P(D) - P(D|\bar{E})]/P(D) = 1 - [P(D|\bar{E})/P(D)]$. Now,

$$\begin{aligned} P(D|\bar{E}) &= P(D|\bar{E})(P(\bar{E}) + P(E)) \\ &= P(\bar{E})P(D|\bar{E}) + \frac{P(D|E)P(E)}{RR}, \end{aligned}$$

The last step following from the definition of *RR*. Hence,

$$\begin{aligned} AR &= 1 - \frac{P(\bar{E})P(D|\bar{E})}{P(D)} - \frac{P(D|E)P(E)}{RR \times P(D)} \\ &= 1 - P(\bar{E}|D) - \frac{P(E|D)}{RR}, \end{aligned}$$

using Bayes' formula twice (see Section 3.4). It follows that

$$AR = P(E|D)\left(1 - \frac{1}{RR}\right). \qquad (5.1)$$

From case-control data we can estimate *P(E|D)* directly, and then with the rare disease assumption estimate *RR* approximately by the estimate of the Odds Ratio. With the data of Table 5.4, this approach yields

$$\widehat{AR} = \frac{50}{100}\left(1 - \frac{1}{2.57}\right) = 0.31,$$

which is very similar to the estimate obtained from the population-based data of Table 5.2. Note that use of the rare disease assumption is questionable here since data from the population-based study reveal that $\hat{P}$(low birth weight infant|unmarried mother)= 0.119, and $\hat{P}$(low birth weight infant|married mother)=0.050, suggesting that *OR* may be substantially larger than *RR;* in fact the estimates from either Table 5.2 or Table 5.3 show that *OR* is approximately 8% greater than *RR*. Again, use of the rare disease assumption can be avoided in using Equation 5.1 under a variant of case-control sampling described in Section 5.4.2.

As was hinted in Section 5.3, it is possible to estimate the Relative Risk or Relative Hazard from case-control samples without the rare disease assumption by modifying the sampling scheme for the controls. On the surface, the rare disease assumption appears to preclude situations where either the disease frequency is high or, essentially equivalently, the interval of risk underlying the definition of disease incidence is sufficiently long so that the cumulative incidence over the entire interval is high. One way to evade the issue of high cumulative incidence is to divide the risk interval into smaller subintervals, chosen so that risk levels in each subinterval meet the rare disease assumption, and then to carry out separate case-control studies for each subinterval

of risk. Odds Ratios can be calculated for each subinterval, and the possibility that these vary over time can be incorporated into the subsequent statistical analysis. In practice, it is natural to implement case-control designs in this fashion when cases can only be sampled as they accumulate in a population. With this in mind, we describe the two most useful and widely used variants of the case-control scheme, the first of which uses exactly the general strategy we just outlined. These modified designs are called *nested case-control studies* because they can be viewed as taking a subsample from a conceptual larger cohort or population.

5.4.1 Risk-set sampling of controls

In a case-control design with risk-set sampling of controls, it is common to select all cases that occur in a population in the defined risk period [0, *T*], although it is perfectly acceptable for only a random sample to be chosen. For each incident case that is identified and sampled at time *t,* one or more controls are randomly drawn from the population of individuals still at risk of disease at *t*. Exposure measurements are taken for each case and for its corresponding set of sampled controls. In essence, this is a stratified, or matched, case-control design where the strata are defined by the times at risk over the interval [0, *T*]. There is no point in sampling controls at times where no disease occurs since they would have no comparative case group. This form of control sampling is widely referred to as *risk set sampling* or *density sampling*. Note that, unlike the traditional or classic case-control sampling of Section 5.3, it is

possible for a control sampled at time *t* to later become a case and enter the sample a second time. Although this is unlikely in large populations unless the disease is common, such a participant must be included in the data set twice. Similarly, it is also theoretically possible that the same individual be selected as a control more than once at differing times, with the same admonition.

Further, note that risk-set sampling of controls accommodates the possibility that the study population is *dynamic* in that individuals may enter and leave the population during the risk interval. The key to the definition of controls in this situation is that they must be *at risk* of being a case (and thus being sampled through that path) at the time of sampling. In addition, the possibility that an individual's exposure level changes over time is also allowed, with the provision that exposure assessment applies *at the time of sampling* for both cases and controls.

To illustrate this form of control sampling, consider a study of the role of the herpes simplex virus type 2 (HSV-2) on the risk for cervical cancer (Lehtinen et al., 2002). In this investigation, the study population consisted of 550,000 women who had donated blood samples to population-based serum banks in Finland, Norway, and Sweden. These samples were collected for a variety of reasons, including first-trimester screening samples during pregnancy and samples from routine health examinations and health promotion projects. Cases of cervical cancer from this study population were identified, over an appropriate calendar risk period, from cancer registries with subsequent linkage to the serum bank using unique identifiers. For each case found, three controls were also chosen from the serum bank that was cancer-free *at the time of diagnosis of the case*. (Controls were also matched to cases by age, geographic subgroups of the bank, and length of storage time for the blood sample, but we defer discussion of this form of matching until Chapter 16.) The donated blood allowed assessment of prior infection with HSV-2 via identification of antibodies. In this case, one possible weakness of the study is that exposure assessment refers to infection status at the time of the donation rather than at the time of sampling. Improving exposure information to avoid this complication would have involved tracing all cases and controls for further blood testing.

With this variant of control sampling, we can calculate expected counts of exposed and unexposed participants in both case and control samples. At any time *t,* let $N_E(t)$ and $N_{\bar{E}}(t)$ be the number of individuals still at risk in the exposed and unexposed population, respectively; the dependence of these numbers on time allows for dynamic changes in the population. The number of cases expected in the exposed group at time *t* is then $N_E(t) \times h_E(t)$, by the definition of the hazard rate, $h_E(t)$, under exposure. Similarly, the number of cases expected in the unexposed group at the same time is $N_{\bar{E}}(t) \times h_{\bar{E}}(t)$, with the same notation. On the other hand, if *m* controls are sampled at *t* in the manner described, then the expected number of exposed and unexposed controls is simply $m \times N_E(t)/[N_E(t) + N_{\bar{E}}(t)]$, and $m \times N_{\bar{E}}(t)/[N_E(t) + N_{\bar{E}}(t)]$, respectively, since the proportion of exposed individuals at risk at time *t* is simply $N_E(t)/[N_E(t) + N_{\bar{E}}(t)]$. Table 5.5 shows these expected counts for each cell in the resulting 2×2 table at time *t*.

The Odds Ratio for this expected data table is quickly seen to be $h_E(t)/h_{\bar{E}}(t) = RH(t)$. If we assume proportional hazards, then $RH(t)$ does not depend on *t,* and so the Odds Ratios for each of these tables at differing times is a constant, equal to *RH*. In this way, a constant Relative Hazard can be derived from a case-control design with risk-set sampling, and can be

estimated using methods for combining Odds Ratios across many 2×2 tables discussed later in the book, in particular in Chapters 9, 16, and 17.

Why does the (sample or data) Odds Ratio from Table 5.5 not estimate the population Odds Ratio? The answer is that, with risk-set sampling, the exposure distribution of the sampled controls does not reflect the exposure distribution of $\bar{D}$s as it does for the classic case-control sampling of Section 5.3. Happily, the distortion introduced by this form of sampling leads the sample Odds Ratio to estimate the Relative Hazard, arguably a more interpretable measure of association.

An important variation of density sampling involves selecting all cases as for risk-set sampling, but sampling controls throughout the risk interval [0, *T*] without regard to the timing of the incident cases. The expected cell entries from the table that pools the cases and controls collected in this way are given in Table 5.6, where we have taken the liberty of using integral signs to reflect that we are pooling the case and control observations over time. If preferred, these integrals can be loosely interpreted as "sums" over distinct short time periods that span the entire risk interval. Note that we allow for the possibility that the number of sampled controls, *m(t),* varies over the risk interval. The Odds Ratio from this table is complex, but can be simplified enormously if we assume that the Relative Hazard, *RH(t),* is constant over time (see Equation 4.5) and that the proportion of exposed individuals, $N_E(t)/[N_E(t) + N_{\bar{E}}(t)] = p_E$, also does not vary with time. In this scenario, the Odds Ratio from Table 5.6 is then

$$OR = [RH \times \int p_E N(t) h_{\bar{E}}(t)\,dt \times \int (1 - p_E) m(t)\,dt]/[\int (1 - p_E) N(t) h_{\bar{E}}(t)\,dt \times \int p_E m(t)\,dt] = RH,$$

where $N(t) = N_E(t) + N_{\bar{E}}(t)$ is the total number at risk at time *t*. Thus again, the sample Odds Ratio, this time from a table pooled over time, estimates the Relative Hazard, albeit with the crucial assumption that the population prevalence of exposure remains constant over time even when the population size varies. Avoiding this assumption by stratifying the pooled table over the time at risk (or, equivalently, time of sampling) takes us back to a form of risk-set sampling, albeit without a fixed number of controls per case in each stratum.

Notice that we did not require any assumption about the frequency with which *D* occurs over the entire risk period (e.g., a rare disease assumption), under either form of density sampling, in showing that a sample Odds Ratio can be seen as an estimate of a constant Relative Hazard.

5.4.2 Case-cohort studies

Suppose n_D cases are selected as with risk-set sampling, or the traditional design for that matter. Now, *m* controls are chosen at random from the entire population at risk at *t*=0 (of size *N,* say), the beginning of the risk period. Exposures are calculated for all sampled participants as usual. Here, the controls may, in fact, include a case and vice versa, so that the word control here means something slightly different from its usage in the traditional design; the sampled controls are often referred to as a sub-cohort of the original Study Population. The sampling scheme is known as the *case-cohort* design.

The Women's Health Trial used a case-cohort approach as part of a general investigation of the effects of a low fat diet on women's health, with particular interest in the risk of breast cancer. In this study, women were randomly assigned to a low fat intervention or control (no

major intervention) group. At 2-year intervals, participants filled out 4-day food frequency questionnaires and blood samples were drawn and stored. Evaluation of disease incidence involved 10 years of follow-up. Assessment of the intervention depended on the full set of enrolled women, but investigation of the role of actual dietary information and blood lipid analyses used the case-cohort approach on a subgroup of 32,000 women of ages between 45 and 69, whose fat intake was high at entry and who possessed at least one known risk factor for breast cancer. In particular, all breast cancer cases were sampled, together with 10% of the original 32,000 as a sub cohort. The case-cohort design minimized the expense of abstraction of the food diaries and laboratory tests. Self et al. (1988) provide a complete description of the study design.

Under case-cohort sampling, the expected total number of exposed (unexposed) controls in the sub-cohort is simply $mP(E)$ ($mP(\bar{E})$, respectively). On the other hand, the expected number of exposed (unexposed) cases is $n_D P(E|D)$ ($n_D P(\bar{E}|D)$). Table 5.7 gives the expected cell entries for the entire sample.

The data Odds Ratio for this table is $P(E|D)P(\bar{E})/P(E)P(\bar{E}|D) = P(D|E)/P(D|\bar{E}) = RR$, so that with this sampling scheme and control definition, the data Odds Ratio actually estimates the population Relative Risk. As for risk-set sampling, we often summarize by saying that in a case-cohort study the data Odds Ratio estimates the Relative Risk, again with no assumption of disease rarity. Similarly to risk-set sampling, the (data) Odds Ratio from Table 5.7 does not estimate the population Odds Ratio because, again, the exposure distribution of the sampled controls (i.e., the cohort) fails to reflect the exposure distribution of $\bar{D}$s. In fact, it yields the total population exposure distribution as can be seen in the right-hand column of Table 5.7. Nevertheless, the distortion introduced by case-cohort sampling leads the sample Odds Ratio to estimate the population Relative Risk. If we prefer to estimate the population Odds Ratio, we can always "remove" any cases from the cohort sample, so that then the right-hand column of Table 5.7 reverts to being a sample of disease-free $(\bar{D})$ controls, in which case the data now have the same structure as a traditional case-control design providing an Odds Ratio estimate as in Section 5.3.

In sum, cases are sampled in identical fashions in all three case-control design strategies, but the subtle differences in the way "controls" are sampled lead the sample Odds Ratio to estimate (1) the population Odds Ratio for classic case-control designs, (2) the Relative Hazard for risk-set sampling, and (3) the Relative Risk for case-cohort studies. This is but one sign that data analysis techniques and their interpretation depend in important ways on how the sample is selected from the Study Population.

That is, the design matters! we will see repeatedly how the various case-control designs impact how we use the sample to both estimate a measure of association and the uncertainty surrounding such estimates.

Conclusion

The case-control design has traditionally been thought to suffer from increased exposure measurement error and selection bias through an inappropriate choice of a control population. However, modern case-control studies are usually designed with careful consideration of these potential problems. Further, the cohort study is also subject to potentially greater error in disease measurement since it often requires long periods of follow-up. Selection bias is also an issue if the exposure groups are not carefully defined. Forms of selection bias in either design are discussed at length in Kleinbaum et al. (1982, Chapter 11). We look at statistical reasons to prefer one of the three basic design strategies in the next chapter.

Case-cohort and nested case-control designs are particularly appealing when general exposure information is collected in a preliminary fashion for all sampled individuals in a large cohort—for example, serum samples or extensive diet histories—but exact exposure measurement from such sources is expensive. Other forms of density sampling are discussed in Langholz and Goldstein (1996). The case-cohort design is well suited to studies of multiple outcomes since the same control sample can be used for each comparison. Wacholder (1991) discusses practical issues concerning the choice of a nested case-control or case-cohort design. Rodrigues and Kirkwood (1990) give a very readable description of the various case-control designs, with practical suggestions for making a specific choice depending on the frequency and acuteness of the disease.There is substantial literature on the appropriate choice of controls and methods for sampling in case-control designs. See Wacholder et al. (1992a,b,c) for an overview of control selection. Random digit dialing (Waksberg, 1978) is often used for control sampling when no convenient sampling frame is available, although this method is becoming increasingly problematic as non-response rates for telephone surveys have raised substantially in recent years. Choosing controls, whether in a cohort or case-control study, rises again the issue of how “representative” the Study Population is of the Target Population with regard to the validity of extrapolation of particular sample information to the Target Population. While “representativeness” is necessary in describing many aspects of the Target Population, it is often not required and may not even be desirable if our sole intent is estimation of a measure of association. For example, in case-control studies, we may have access to an accurate registry of cases for a well-defined subset—restricted, for example, by geography or the nature of the cases—that is not representative of all cases in the population. In this case, we are still able to implement a successful case-control design for estimating the Odds Ratio, say, by ensuring that disease-free individuals, or controls, are selected at random from a Study Population that adequately serves as the source population only for the restricted group of cases.

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.