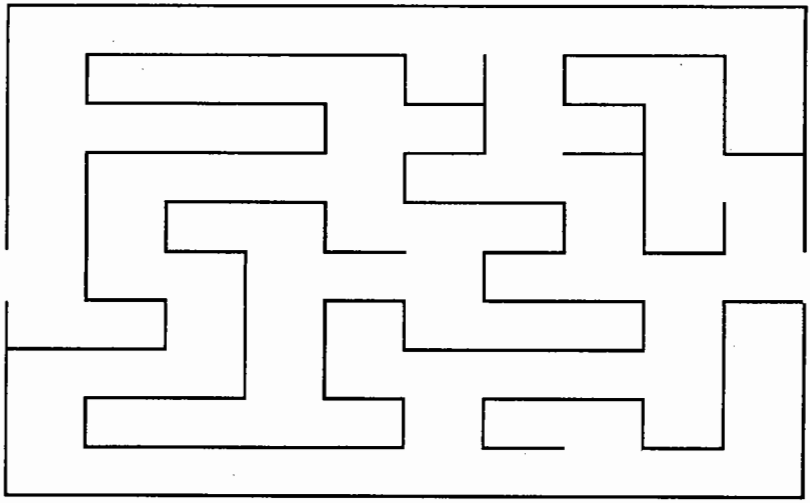
7

Generalization



The purpose of this chapter is to introduce the vocabulary used to link theory and observations. Parameters are theoretical values; they determine the shape and location of the long-range frequency distributions for observed data. Two parameters will be discussed: probability, which characterizes both prevalences and cumulative incidences, and hazard, which characterizes incidence rates. Both statistical and systematic errors in parameter estimates will be addressed. Confounding is an example of systematic error that arises when a planned comparison juxtaposes noncomparable groups. There is also the uncomfortable, empirical fact that epidemiologic observations are less fully replicable than statistical models suggest they ought to be.

Parameters

While rare exposures, such as standing at ground zero of a hydrogen blast, may absolutely determine vital status, most determinants of health are accompanied by an element of uncertainty. Even very similar individuals (identical twins, for example), do not have precisely the same health status throughout their lives.

A deterministic view of disease causation can account for uncertainty about the cumulative incidence for any particular group by holding that study groups are the imperfect reflection of an underlying reality. The "reality" is envisioned as the class of all possible subjects meeting study criteria, and the studied property of that infinite class is the proportion destined, for example, to have a heart attack within a year. If the presence of that characteristic is unknown on an individual level, then the proportion that characterizes the class can differ from the observed cumulative incidence because of vagaries in the selection of the study population. The class feature that the observed data approximate, the overall proportion in the imagined class of similar individuals, is referred to as the *probability* of disease.

A probabilistic view of disease causation postulates an unknown mechanism operating within individuals, a black box, that behaves as if governed by a random process with stable long-term characteristics. The long-term fraction of persons who manifest disease is the probability that characterizes the process. Chance here is only a metaphor for the true mechanism of disease production, just as it is only a metaphor for the process of subject selection from an imaginary universe of all possible subjects in the preceding view.

For the purposes of inference from observation to the underlying reality, the arguments presented later will assume that there is a general characteristic of the process giving rise to disease in a particular study group that can be meaningfully summarized as a single number between zero and one, inclusive.

Probability *is a characteristic of the physical processes that give rise to observable events, and represents the limiting value that would be observed for a cumulative incidence or a prevalence as larger and larger numbers of individuals came under scrutiny.*[52]

52. "Risk" is a synonym for probability.

The true probability of an event in an individual is either one or zero for a determinist, and the 20 percent that might be estimated from group data is a measure of the determinist's degree of belief that any one of the individuals who comprise the group actually possesses sufficient elements for the manifestation of disease. For a probabilist, who accepts the black box, the probability of disease in an individual is the expected value of the cumulative incidence in essentially similar persons.

The probability of an event is a *parameter*, the procedure for obtaining a cumulative incidence is an *estimator*, and the cumulative incidence calculated from a particular set of data is an *estimate* of the probability.

Parameter. *The terms other than those describing the circumstances of observation and the outcome in the formulaic presentation of a probability distribution are parameters. Parameters are not observable, but may be estimated from observations.*

Estimator. *An estimator is a procedure for obtaining estimates. It is, equivalently, a random variable whose realization, the "estimate," will be taken as a measure of a parameter. The estimator is a function of random variables whose realizations are the data points being observed.*

Estimate. *An estimate is a realization of the estimator. The estimate is a function of the observed data.*

The following section will address the ways in which estimates provide information about parameters.

Probability serves the epidemiologist as the parameter corresponding to the prevalence of a characteristic or the cumulative incidence of disease; for the statistician, probability plays another role. In a hypothetical population of repeated identical studies, probability also describes the proportion of studies expected to have a particular outcome. "Statistical Uncertainty" (below) will speak to this aspect of probability.

In order to make the transition from the probability parameter to a parameter that corresponds to the incidence rate, we need to adopt some formalisms for the description of time and temporal relations. A bracket, "[" or "]", next to a time designation means that a starting or stopping time is included in the interval that it bounds. An ordinary parenthesis, "(" or ")", means that the time point is not

included in the interval in question. Written symbolically, the interval that begins just after t_1 and continues through time t_2 is $(t_1,t_2]$, and the cumulative incidence, CI, over that interval is $CI(t_1,t_2]$. In the absence of loss to follow-up during $(t_1,t_2]$, the defining equation for the cumulative incidence is

$$CI(t_1,t_2] = \frac{\text{Cases}(t_1,t_2]}{N(t_1)}$$

$N(t_1)$ is the size of the population that is at risk to become an incident case at time t_1. Denote the probability of acquiring disease during the interval $(t_1,t_2]$ by $R(t_1,t_2]$. CI is an estimate of R. Clearly, as t_2 gets closer to t_1, $R(t_1,t_2]$ approaches zero. However, by dividing $R(t_1,t_2]$ by the length of the time interval over which the cumulative incidence is calculated (obtained by subtracting t_1 from t_2), it is possible to obtain a stable, limiting value, characteristic of t_1, called the *hazard*, and denoted symbolically by h.

$$h(t_1) = \lim_{t_2 \to t_1} \frac{R(t_1,t_2]}{t_2 - t_1}$$

$$= \left. \frac{dR(t_1,t_2]}{dt_2} \right|_{t_2=t_1}$$

The symbol "$\lim_{t_2 \to t_1}$" means "the value approached as t_2 comes closer to t_1." The convention in the second line of the above equation describes incremental changes; "dx/dy" means "the incremental change in x associated with each change in y." The vertical bar with the subscript "$t_2=t_1$" at the end of the expression means "evaluated when t_2 equals t_1," that is right at the beginning of follow-up. The second line thus refers to the incremental change in the probability of survival immediately following t_1. The incremental changes could be read as the slope at $t_2=t_1$ of a graph of $R(t_1,t_2]$ versus t_2.[53]

Hazard. *The hazard is the limiting value of the probability of becoming an incident case per unit time among those at risk for becoming a case.*

53. Look again at the daily incidence curve of Figure 1.3.

The hazard function has the units "change in the number of cases per population size (N) per unit time," so that the unit of measurement, or *dimension*, of a hazard is "cases per *person time*." The units of the daily incidence of bleeding shown in Figure 1.3 would be cases per person day.

The dimension of hazard as it appears in the literature of survival theory, and in some epidemiologic texts, is the reciprocal of time. The transition from "cases per person time" to "per time" is achieved by observing that "cases" are simply a count and therefore dimensionless, and that "person time" is a cumulation over persons of times observed in those persons, and therefore has the dimension of "time."

Defined as they have been above, hazards characterize instants. During short intervals, an insufficient number of events occurs to provide a useful estimate of hazard. Therefore, observations designed to permit an estimate of the hazard invoke the assumption that there are periods of time and population definitions that together can specify some finite quantity of person time during which, to a reasonable approximation, the hazard can be taken to be constant. Within such blocks of person time, an estimate of the hazard is provided by the incidence rate.

The relation between incidence rate and hazard is analogous to that between cumulative incidence and probability. The incidence rate is the observable counterpart of hazard. The hazard is the parameter of which the incidence rate is an estimator. The hazard is the value to which the incidence rate would tend as the amount of person time under observation became larger and larger. Refer again to Table 1.2.

Statistical Uncertainty

When the observed data are precisely those that would be expected under a hypothesis that some particular parameter value is true, then we take that hypothesis as an estimate of the parameter. The concept of statistical uncertainty permits an extension of this procedure to statements about the consistency of data with parameters that predict something else, and even to situations in which the data observed would not have been expected under any single parameter. (This arises when different estimates of a single quantity disagree.) A good estimate will be a value that, if it were the parameter, would

place the observed data in a position of highest possible probability. There is, however, no guarantee that such an estimate actually equals the estimated parameter.

The unknown distance that separates an estimate from its corresponding parameter raises the problem of statistical inference. Uncertain observations are compatible with an infinite number of parametric "truths," and one job of the statistician is to lay down limits on the kinds of realities that might have given rise to a set of data.

Figure 7.1 presents an idealized picture of the relative frequency of different estimates of a parameter whose true value has been taken for the purposes of illustration as 5. The curve is an example of a *probability density function*, and has a highly characteristic shape. The area under the curve is precisely one unit, and the probability that any particular estimate will fall between two values is equal to the area under the portion of the curve bounded by those values.

From inspection of Figure 7.1, it is evident that the greatest probabilities are associated with estimates near the true parameter value, and that the probability of estimates in an interval of any particular size falls off rapidly as the interval becomes removed from the true parameter value. The whole open-ended interval that begins 1.96 units above the parameter has an area of 0.025, indicating that such extreme values occur in about 2.5 percent of estimates. The curve is symmetrical, and the sum of the two tail areas that are bounded by 1.96 units above and below is five percent.

The units of the horizontal axis of Figure 7.1 are *standard errors*, and the shape of the curve is given by the so called *Normal* or *Gaussian* probability distribution. The curve, first described by Carl Friedrich Gauss to account for errors of measurement in astronomy, is the limiting form of the error distribution of all epidemiologic measures.

Normal distribution, *also called the Gaussian distribution, is the probability density function that describes the distribution of realizations x of a continuous random variable X when the value x is the sum of a very large number of random variables whose probability distribution is arbitrary, but whose variances (see below) are of similar magnitude.*

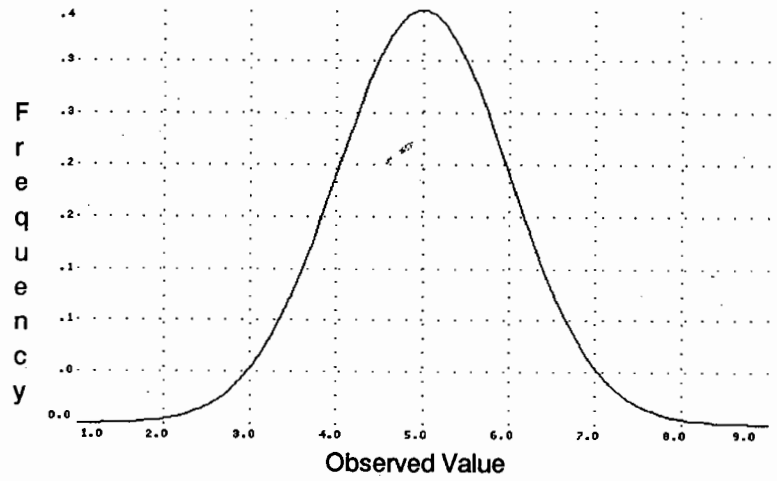


Figure 7.1 Distribution of estimates of a hypothetical parameter

The standard error is the square root of the *variance*, a measure of the dispersion of a probability distribution. In order to assess the large sample properties of any estimator of a parameter, the analyst needs only the parameter value and the standard error of the estimator.

Standard error. *The standard error of an estimate is the square root of the variance of the estimator.*

Variance. *The variance of a random variable is the expected value of the square of the deviation of x from the expected value of X.*[54]

By one convention, estimates that fall more than 1.96 standard errors away from the parameter value are considered improbable (though they do occur five percent of the time). The importance of the convention is that it opens a path for inductive reasoning: begin with the observations and proceed to a statement about an unknown parameter. The trick is to hypothesize a parameter value, calculate the probability distribution of estimates under the hypothesis, and

54. See below and Chapter 13 for formal definitions of "expected value."

then to locate the known estimate on that distribution. If the estimate, which is known, would be improbable under the hypothesis represented by a particular parameter value, then the parameter value is *rejected* as a candidate explanation for the data.

Two widely used applications of this reasoning process are the *p value* and the *confidence interval*. The p (for probability) value is the tail area beyond the given estimate, assuming a particular parameter. The parameter assumed is most commonly one that represents a nil effect, and the p value is the probability of getting either the estimate actually obtained or an estimate further from the nil value. If this tail probability is low, and the hypothesis of a nil effect is rejected, then it follows that the effect must be non-nil. When the p value is less than some prespecified cutoff, such as five percent or one percent, then the estimate differs from the nil effect level by a *statistically significant* amount.

p value. *The p value is the probability of occurrence of estimates that are as or more deviant from posited parameter values than the estimates actually obtained from a body of data. The p value is a function of observed data. It is the realization of a random variable whose distribution is uniform in the range* [0, 1] *under posited parameter values, and whose distribution becomes non-uniform, with an increased density near zero, under specified kinds of deviation from the posited values.*

Confidence intervals provide a more exhaustive application of statistical inference. A range of acceptable parameter estimates can be derived as those for which the observed data would not be too improbable. For any estimate, two hypothetical parameter values, one below the estimate and the other above, are identified to meet the following criterion: the size of the tail area that the estimate cuts off from the probability distributions implied by each of the two hypothetical parameter values is some prespecified amount. The sum of the two tail areas conventionally is either five or ten percent. If the tail areas sum to five percent, then the interval between the hypothetical parameters is a *95 percent confidence interval*; if the tail areas sum to ten percent, then the interval is a 90 percent confidence interval.

Confidence interval. *A confidence interval is a set of possible parameter values that are consistent with a body of observations in the sense that the p values for the data given any of the parameter values in the interval are greater than a specified amount, usually designated by* α. *The salient operational feature of a confidence interval is that it is calculated by a mechanism that has a priori a* $1-\alpha$ *probability of including the true parameter value.*

For the measures considered in this text, a large-sample 95 percent confidence interval can always be constructed as follows.

(1) Identify the standard error associated with a particular measure. This value is the square root of the variance of the measure, and is a function of the parameter being estimated and of the number and the characteristics of subjects studied.

(2) Calculate the distance separating a parameter from an estimate that corresponds to a tail area of 2.5 percent. The distance is 1.96 times the standard error.

(3) Add the calculated distance to the estimate to obtain the upper 95 percent confidence bound.

(4) Subtract the calculated distance from the estimate to obtain the lower 95 percent confidence bound.

The parameter values that lie within the 95 percent confidence interval constitute a set of possible realities that are consistent with the observed data.

The arbitrariness of a choice of 90 or 95 percent confidence should be evident. The utility of the interval is not explicitly to include or exclude parameter values of interest, but rather to provide an indication of the range of true values that may have given rise to a given set of study results. A number of examples of confidence interval calculations are given in Chapter 8.

Confounding

The analysis of random error presupposes that there is no difference between comparison groups such as might give rise to different disease frequencies, other than the factor that is used to define the groups. Even in a carefully designed study, there is no guarantee of this kind of comparability. The distortion of analytic results that can arise from dissimilar comparison groups is called *confounding*. Confounding produces an *expected value* of the estimate that is different from the value of the parameter being estimates. Confounding is a form of *bias*.

Expected value. *The expected value of a random variable X is the average value that is observed in many repeated realizations of X.*

Confounding. *When imbalances in the composition of compared groups give rise to an expected value of a comparative measure that differs from the effect of the factor that defines the groups, the estimate of the effect of that factor is said to be confounded.*

Bias. *The difference between the expected value of an estimator and the parameter whose value is being estimated is the bias of the estimator.*

The relation of confounding to the characteristics of the study population will be explored in a separate chapter. One proper way to deal with confounding by factors that can be measured is to separate the study groups according to levels of the confounding factor. This is the core of stratified analyses, which will be dealt with in Chapter 8. The anticipated magnitude of confounding in an analysis that ignores the confounding factor is the subject of Chapter 9.

The inclusion of the term "expected value" in the definition of confounding implies the existence of chance mechanisms that could lead to estimates not equal to the parameter value even in the absence of confounding. In a fully deterministic view of disease causation, the "chance" processes are unmeasured determinants of disease, and any net contribution of those factors to disease is a form of confounding.

Uncertainty Missed by Statistical Models

In small studies, the estimated effects of chance may overwhelm errors arising from other sources, and the p values and confidence intervals calculated in standard ways may provide useful guides to the imprecision of estimates. This utility does not extend to large studies. Chance, with its estimable errors, is sadly not the principal source of invalidity in most observational research.

Compilations of estimates from multiple studies are undertaken with increasing frequency, and statistical uncertainty has been found regularly to understate interstudy variability, particularly when the numbers of observations have been sufficiently large as to reduce the statistical uncertainty to modest proportions. For this reason, inference based on statistical considerations alone gives a more

optimistic picture of the precision of knowledge than the data really justify. The chief role for the confidence interval estimates found in Chapter 8 is to set an upper bound to the analyst's certainty about the meaning of the results in hand.

Strict determinism and probabilistic views of disease causation differ strikingly in the directions in which they look to resolve questions of residual uncertainty. The determinist, driven by the idea that the origins of every instance of illness are knowable, will pursue details of exposure and host characteristics that are so individual as to reduce epidemiology to case reports. By contrast, the probabilist's black box can be enlarged to any dimension, and he is disinclined to pursue population differences that he ascribes to chance. Although the latter view empirically seems associated with less time wasted in the pursuit of the unknowable, the former leads to most new understanding of causal relations: the specific well-documented instance becomes the paradigm for a previously unimagined class. In the area of medical statistics, the role of the determinist is most often adopted by the investigating clinician, that of the probabilist by the statistician. Epidemiologists, who may come out of either tradition, do best when they keep a foot in either camp.