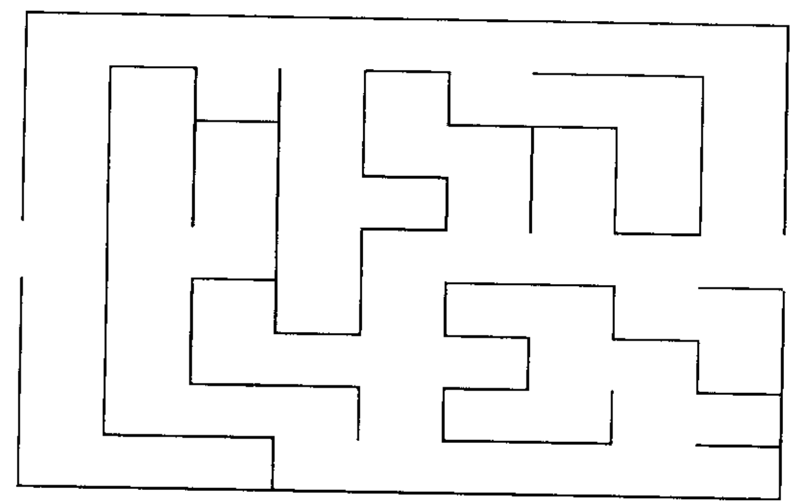
8

Estimation



On the basis of data, we make estimates. The estimates are constructed in such a way that they bear a known relation to the parameters of processes that gave rise to the data. This chapter describes the mechanics of obtaining estimates and of judging the relation between estimate and parameter for closed and open cohort studies, and for case-control studies. All the statistical terms used here are defined in Chapter 13.

Analysis of Closed Cohort Studies

The methods presented here will be illustrated using the data of Example 5.1, which you should review before continuing.

Calculation of attack rates and cumulative incidence differences. Since pharyngitis could only occur once during the period of observation, the designation of each of the cohort members as a case or non-case was unambiguous, and it was reasonable to calculate the fraction of persons in each cohort who became ill. This is the cumulative incidence, as defined in Chapter 1. Infectious disease epidemiology has its own term for the cumulative incidence: *attack rate* (*AR*).[55]

Attack rate. *The attack rate is the cumulative incidence of disease in persons who are exposed to an agent whose effect is shorter than the time of potential follow-up. The period of follow-up begins at the time of exposure and continues over a closed interval that allows the expression of all possible new cases attributable to the exposure.*

The attack rate provides an estimate of the probability of infection that each of the cohort members faced at the beginning of the convention. Since the attack rate and the cumulative incidence are identical quantities, the latter term will be used in the remainder of this section for consistency. Readers should bear in mind that the infectious disease literature uses the former term. The cumulative incidence difference is obtained by subtracting the cumulative incidence in an unexposed group from that in an exposed group. Thus, in Table 5.1, the cumulative incidence difference associated with luncheon attendance is (47/86)-(11/77) or 40.4 percent in those who attended the dance and (8/23)-(1/40) or 32.3 percent in those who did not attend the dance.

Cumulative incidences and cumulative incidence differences observed in particular studies are estimates. Approximate confidence intervals for each of the corresponding probabilities and for their differences can be derived by assuming that the number of cases is distributed as a binomial variable.

55. The word "rate" in this term is a misnomer in the system of nomenclature presented here, since the attack rate is not a rate but a proportion.

Binomial distribution *is the probability distribution that describes the number of events observed in N opportunities to observe an event, when the probability of observing a single event at any opportunity is π, and is unaffected by the observation of an event at any other opportunity.*

$$\Pr(x \mid N) = \frac{N!}{x!(N-x)!}\pi^x(1-\pi)^{N-x}$$

$$E(X) = \pi N$$

$$\mathrm{Var}(X) = N\pi(1-\pi)$$

The range of possible values for x is [0,N]. π is the binomial parameter.

In this definition, as elsewhere, "$\Pr(x|N)$" means "the probability of observing x cases among N observed persons"; "$E(X)$" means "the expected value of the random variable X"; and "$\mathrm{Var}(X)$" means "the variance of the random variable X."

In the calculation of a cumulative incidence, x is the number of cases and N is the number of individuals in the population at risk; the cumulative incidence is given by x/N. The variance of x/N is $x(N-x)/N^3$.[56] The bounds of an approximate 95 percent confidence interval for the probability of pharyngitis, of which x/N is an estimate, are obtained by adding and subtracting 1.96 times the standard error (the square root of the variance) to and from x/N. For those who attended the luncheon and the dance, x is 47 and N is 86, so that

$$CI = AR = x/N = 47/86 = 0.547$$

The 95 percent confidence bounds for the probability of pharyngitis are

56. c.f. Table 1.2 and Chapter 13. Note that

$$CI = \frac{x}{N}$$

$$\frac{CI(1-CI)}{N} = \frac{x(N-x)}{N^3}$$

$$\text{lower} = \frac{x}{N} - 1.96\sqrt{\frac{x(N-x)}{N^3}}$$

$$= \frac{47}{86} - 1.96\sqrt{\frac{47(86-47)}{86^3}}$$

$$= 0.441$$

$$\text{upper} = \frac{x}{N} + 1.96\sqrt{\frac{x(N-x)}{N^3}}$$

$$= \frac{47}{86} + 1.96\sqrt{\frac{47(86-47)}{86^3}}$$

$$= 0.652$$

The cumulative incidence for conventioneers who attended both the luncheon and the dance was 55 percent, with 95 percent confidence bounds of 44 and 65 percent. (In the above calculation, an extra digit has been retained in the values derived in intermediate steps for accuracy. In practice, all intermediate values should be retained with as many digits as possible, with rounding employed only for the final result.)

The calculation of the confidence interval above has two important limitations. First, the method employed is a *large sample* technique, whose accuracy improves as the number of study subjects becomes larger. When the smallest count involved in the calculation is larger than 30, then the results are virtually identical to more accurate methods of calculation, which can be found in intermediate textbooks of epidemiology or statistics. When the smallest count is ten or greater, the results are close enough for most any purpose; when the smallest count is five or less, the method gives bounds that are only roughly indicative of the range of parameter values.

The second important feature is that the method assumes independence of individual results. The risk in any individual is assumed not to be affected by the outcome in other individuals. In contemporary epidemiology, this means that the method given is inappropriate when the number of ill persons is the result of person-to-person transmission of risk.

The variance of the cumulative incidence difference can be estimated by summing the estimated variances of the cumulative incidences that make up the cumulative incidence difference. (See Chapter 13 for rules about the manipulation of variances.) An estimate of the variance associated with the cumulative incidence difference between the dancers attending the luncheon and those not doing so is

$$\text{Var}(CID) = \frac{47(86-47)}{86^3} + \frac{11(77-11)}{77^3}$$

$$= 0.004472$$

Thus

$$CID = \frac{47}{86} - \frac{11}{77}$$

$$= 0.4037$$

and the 95 percent confidence bounds for the difference in the probabilities of pharyngitis are

$$\text{lower} = 0.4037 - 1.96\sqrt{0.004472}$$

$$= 0.273$$

$$\text{upper} = 0.4037 + 1.96\sqrt{0.004472}$$

$$= 0.535$$

Summarizing experience across several strata. The cumulative incidence differences associated with attending the luncheon in those who attended the dance and in those who did not attend the dance are somewhat different (40.4 percent and 32.3 percent, respectively). There are a variety of ways in which these two estimates might be summarized in a single overall figure. One approach is to take a weighted average of dancers' and nondancers' cumulative incidences among the luncheon attendees, and then to do the same for the nonattendees, using the same weights. The weighted averages are said to be *adjusted* for the effects of attendance at the dance. Because the same weighting scheme is used to adjust the rates of the luncheon "exposed" and "unexposed" groups, the comparison between exposure groups is a valid one. The resulting weighted averages are referred to as *standardized* cumulative incidences.

Standardization. *Standardized measures are formed from a series of individual measures by taking a weighted average of the individual values.*

Standard. *The set of weights used for standardization is the standard. These weights sum to 1.*

One convenient standard is the distribution of dancers among luncheon attendees. Of the attendees, 86/(86+23) or 78.9 percent went to the dance, and 23/(86+23), or 21.1 percent, did not. Call the cumulative incidences that are standardized over categories of dance attendance the *standardized cumulative incidences* (*SCI*). The cumulative incidences for dancers and nondancers among luncheon attendees were 47/86 (54.7 percent) and 8/23 (34.8 percent), respectively. The *SCI* for luncheon attendees is

$$SCI(\text{exposed}) = \left(\frac{86}{86+23}\right)\left(\frac{47}{86}\right) + \left(\frac{23}{86+23}\right)\left(\frac{8}{23}\right)$$

$$= 0.5046$$

For those who did not attend the luncheon, the cumulative incidences among dancers and nondancers were 11/77 (14.3 percent) and 1/40 (2.5 percent), respectively. The *SCI* for nonattendees is

$$SCI(\text{unexposed}) = \left(\frac{86}{86+23}\right)\left(\frac{11}{77}\right) + \left(\frac{23}{86+23}\right)\left(\frac{1}{40}\right)$$

$$= 0.1180$$

The standardized cumulative incidence difference (*SCID*) is the difference between the standardized cumulative incidences.

$$SCID = SCI(\text{exposed}) - SCI(\text{unexposed})$$

$$= 0.3866$$

Note that this result is identical to the result that would be obtained by applying the standard weights to the stratum-specific cumulative incidence differences. The cumulative incidence differences are 0.4037 for those who attended the dance and (8/23)-(1/40) = 0.3228 for those who did not. Standardized over categories of dance attendance, the cumulative incidence difference would be

$$SCID = \left(\frac{86}{86+23}\right)(0.4037) + \left(\frac{23}{86+23}\right)(0.3228)$$

$$= 0.3866$$

The standardized cumulative incidence difference can be seen as a weighted average of the component cumulative incidence differences.

The variance of a weighted average is given by the sum of the component variances, each weighted by the square of the corresponding weight, so that

$$\text{Var}(SCID) = \left(\frac{86}{86+23}\right)^2 (0.004472)$$

$$+ \left(\frac{23}{86+23}\right)^2 (0.01047)$$

$$= 0.003250$$

95 percent confidence bounds to the *SCID* are therefore

$$\text{lower} = 0.3866 - 1.96\sqrt{0.003250}$$

$$= 0.2749$$

$$\text{upper} = 0.3866 + 1.96\sqrt{0.003250}$$

$$= 0.4983$$

The luncheon "effect" is standardized according to the dancing choices of those who actually attended the lunch. The final measure is intuitively satisfying in that it is directly tied to the experience of the exposed group: it addresses the question "What would have been the difference between those who attended the luncheon and those who did not if the nonattendees had the same fraction of dancers as those who attended?" Other weighting schemes that might have been used include an external standard (such as equal weights for each group), or an internal standard based on the reciprocals of the variances[57] of the stratum-specific cumulative incidence differences. This last standard minimizes the variance of the final

57. The reciprocal of the variance of an estimate is also known as the "information" contained in the estimate. Very precise estimates have small variance and high information.

estimate, but its proper use presupposes that the only source of discrepancy between the stratum-specific cumulative incidence differences is chance.

Confounding. If the interest in Table 5.1 had focussed on the cumulative incidence difference associated with attendance at the dance, the investigators could have calculated estimates of (47/86) - (8/23) or 20 percent in those who attended the luncheon and (11/77) - (1/40) or 12 percent in those who did not. Any standardized estimate of an overall effect would lie between these two values. If luncheon attendance were ignored, a crude cumulative incidence difference might also have been calculated as

$$CID = \frac{47+11}{86+77} - \frac{8+1}{23+40}$$

$$= 0.213$$

or 21 percent. This value lies outside of the range of stratum-specific estimates. Because luncheon attendance was more common among dancers than among those who did not dance, the crude cumulative incidence difference reflects a part of the cumulative incidence associated with luncheon attendance, in addition to the effect of dance attendance on risk. The crude cumulative incidence difference therefore provides a biased estimate of the increase in probability of pharyngitis associated with attendance at the dance.

Analysis of Open Cohort Studies

Example 5.2 will be used to illustrate the techniques presented here.

Error estimates and comparisons of incidence rates. Just as the observed proportion of the disease in a closed cohort study is an estimate of the underlying probability of developing disease, so the ratio of cases to person time, the incidence rate, provides an estimate of the underlying hazard of disease. The most straightforward technique for assessing the variability of incidence rates in open cohort studies is based on a treatment of the incidence rate calculation as if the numerator (the number of cases) were variable and the denominator (the amount of person time) were fixed. If x is the

number of observed events and P is the person time at risk, then x is the realization of what is called a Poisson process. The probability distribution from which x is drawn is the Poisson distribution.[58]

Poisson distribution *is the probability distribution that describes the number of events observed in a block of person time when the expected number of events is directly proportional to the total person time of observation. Let θ be the expected number of events per unit of person time and $\lambda = \theta P$ be the number of events expected in a block of person time of size P.*

$$\Pr(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

$$E(X) = \lambda$$

$$\mathrm{Var}(X) = \lambda$$

The range of possible values for x is $[0, \infty)$. λ is the Poisson parameter. If P is imagined as being composed of a very large number of discrete units of person time, so that the probability of an event in any person time unit is very small, then the probability distribution of the number of events in P may also be considered to be binomial, with N taken as the number of discrete person time units. All the formulas above are derivable from their binomial counterparts in the limiting case in which N approaches infinity, with P and λ constant.

The number of observed events x is an estimate of a Poisson parameter λ. The incidence rate estimate IR is given by x/P, with variance x/P^2.[59] The mortality rate estimate and its variance for the period from 30 through 34 years since first exposure (Table 5.2) are given by

---

58. The development here presumes that the expected number of cases is directly proportional to the amount of person time of observation. Put another way, we presume that there is no element of contagion, in which the probability of a case occurring is a function of the number of other cases that have occurred.

59. c.f. Table 1.2 and Chapter 13. Note that

$$IR = \frac{x}{P}$$

$$\frac{IR}{P} \bullet \frac{x}{P^2}$$