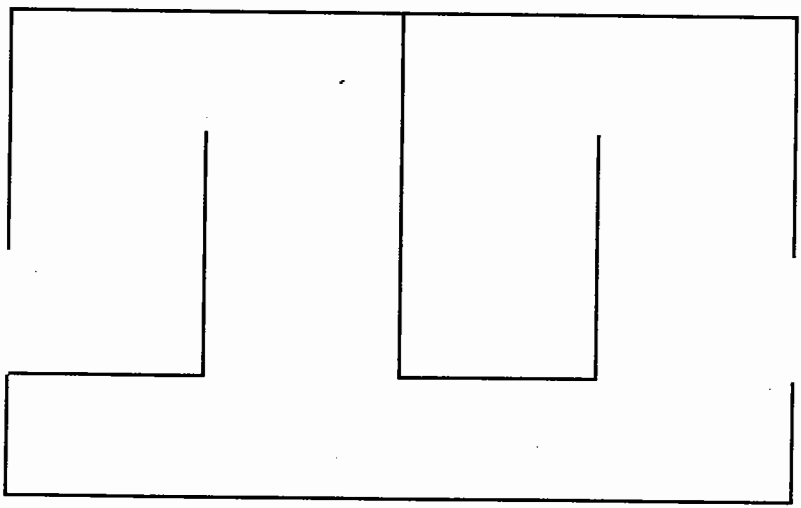
13

Statistical Terms and Concepts



Although the concepts of contemporary epidemiology are not mathematical ones, there is no useful manipulation of the epidemiologic ideas that does not resort at some point to statistical inference. In this sense, epidemiology depends upon statistics in much the same way that physics depends on mathematics. What follows is not intended to substitute for a formal introduction to applied statistics, but is rather a sketch of a number of statistical terms used in epidemiology. The first section lists definitions of commonly used statistical terms; the second provides the rules by which the variance estimates appropriate to different sampling distributions can be combined to provide variance estimates for epidemiologic measures.

Definitions

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

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.

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.*

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

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.*

Expected value. *The expected value of a random variable X is the average value that is observed in many repeated realizations of X. The expected value can be calculated from the probability distribution as*

$$E(X) = \sum x \Pr(x)$$

where the summation is over all possible values of x. It can be obtained from the probability density function as

$$E(X) = \int x f(x) dx$$

where the integration is over all possible values of x. The expected value is a measure of the location of the probability distribution in the universe of possible values of x.

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.*

Mean square error. *The mean square error of an estimate is the expected value of the square of the deviation of the estimate from the parameter value of which it is an estimate. The mean square error can be calculated from the probability distribution as*

$$\mathrm{MSE}(\hat{\mu}) = \sum (\hat{\mu} - \mu)^2 \Pr(\hat{\mu})$$

and from the probability density function as

$$\mathrm{MSE}(\hat{\mu}) = \int (\hat{\mu} - \mu)^2 f(\hat{\mu}) dx$$

where μ is the parameter and the symbol " ^ " indicates an estimate of the parameter. Mean square error is a measure of the dispersion of the realizations of the estimate around the parameter value. It is the sum of the variance and the square of the bias.

$$\mathrm{MSE}(\hat{\mu}) = \mathrm{Var}(\hat{\mu}) + [\mathrm{Bias}(\hat{\mu})]^2$$

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 are of similar magnitude.*

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$E(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$

The range of possible values for x is (-∞, +∞). μ and σ², the mean and variance, are the parameters of the Normal distribution. For the binomial and Poisson distributions, when the distribution of X is such that the probability of a realization at or near a limiting value is nearly zero, many of their properties can be approximated by considering them to be Normal distributions whose expected values and variances are derived from the corresponding binomial and Poisson definitions.

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.*

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.*

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 λ = θP 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, ∞). 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.

Probability. *The probability of observing a particular realization x of a random variable X is the fraction of instances in which x will be observed out of many repeated observations of X.*[70]

Probability density. *The probability density is the counterpart of the probability distribution for continuous random variables. The probability density is always described in formulaic terms as the "probability density function," designated by f(x). The probability of observing a realization x of X in the range [a,b] is*

$$\Pr(X \in [a,b]) = \int_a^b f(x)dx$$

Probability distribution. *The probability distribution of a random variable X is the collection of the probabilities of all possible realizations of X. A probability distribution can be characterized by listing or graphing the relation between x and Pr(x), or by presenting a formula that permits the calculation of Pr(x) for any value of x. The terms in such a formula other than x and terms describing the circumstances of observation are called "parameters." Probability distributions may also be characterized approximately by presenting their expected values and variances.*

Standard deviation. *The standard deviation of a random variable X is the square root of Var(X).*

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. The variance can be calculated from the probability distribution as*

$$\mathrm{Var}(X) = \sum [x - E(X)]^2 \Pr(x)$$

and from the probability density function as

70. See also Chapter 14.

$$\mathrm{Var}(X) = \int [x - \mathrm{E}(X)]^2 f(x)\,dx$$

Variance is a measure of the dispersion of the realizations of X around the expected value.

Manipulating Variances

Most epidemiologic measures are composites of random variables and functions of random variables. In order to derive the variances of those measures, it is generally sufficient to apply one or both of the following rules.

Variance of a sum

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\mathrm{Cov}(X_1, X_2)$$

Variance of a function

$$\mathrm{Var}[g(X)] = \left(\frac{\partial g}{\partial X}\right)^2 \mathrm{Var}(X)$$

where $\mathrm{Cov}(X_1,X_2)$ is the covariance of X_1 and X_2. In the development below, the epidemiologic measures being summed will be taken to be independent, so that $\mathrm{Cov}(X_1,X_2)=0$.

Table 13.1 Some useful derivatives

Function	*Derivative*
kX	k
$\ln(X)$	X^{-1}

If X is binomial and g(X) = X/N, that is to say if g(X) is a risk

$$\mathrm{Var}\left(\frac{X}{N}\right) = \left(\frac{1}{N}\right)^2 N\pi(1-\pi)$$

$$= \left(\frac{1}{N}\right)^2 \frac{N\pi(N - N\pi)}{N}$$

which can be estimated by substituting x for $N\pi$

$$\mathrm{Var}\left(\frac{X}{N}\right) \doteq \frac{x(N-x)}{N^3}$$

If X is Poisson and g(X) = X/P, that is to say if g(X) is an incidence rate,

$$\mathrm{Var}\left(\frac{X}{P}\right) \doteq \left(\frac{1}{P}\right)^2 \lambda$$

which can be estimated by substituting x for λ

$$\mathrm{Var}\left(\frac{X}{P}\right) = \frac{x}{P^2}$$

The variance of the natural logarithm of an incidence rate (the estimate of a hazard) is

$$\mathrm{Var}\left[\ln\left(\frac{X}{P}\right)\right] = \left(\frac{\lambda}{P}\right)^{-2} \frac{\lambda}{P^2}$$

which can be estimated as

$$\mathrm{Var}\left[\ln\left(\frac{X}{P}\right)\right] \doteq \frac{1}{x}$$

The variance of a rate difference (the estimate of a hazard difference) is

$$\mathrm{Var}\left(\frac{X_1}{P_1} - \frac{X_0}{P_0}\right) = \frac{\lambda_1}{P_1^2} + \frac{\lambda_0}{P_0^2}$$

which can be estimated as

$$\mathrm{Var}\left(\frac{X_1}{P_1} - \frac{X_0}{P_0}\right) \doteq \frac{x_1}{P_1^2} + \frac{x_0}{P_0^2}$$

The variance of the logarithm of a rate ratio (the estimate of a hazard ratio) is

$$\mathrm{Var}\left[\ln\left(\frac{X_1/P_1}{X_0/P_0}\right)\right] = \mathrm{Var}\left[\ln\left(\frac{X_1}{P_1}\right) - \ln\left(\frac{X_0}{P_0}\right)\right]$$

$$= \frac{1}{\lambda_1} + \frac{1}{\lambda_0}$$

which can be estimated as

$$\mathrm{Var}\left[\ln\left(\frac{X_1/P_1}{X_0/P_0}\right)\right] \doteq \frac{1}{x_1} + \frac{1}{x_0}$$