1

Population Proportion or Prevalence

To quantify the impact of a given disease on public health in a community, or in studying the variation of a disease distribution between geographical regions to locate the potential causes, we may wish to first estimate the prevalence of the disease, defined as the population proportion of subjects who have it. In this chapter, we start by discussing the estimation of population prevalence under the most commonly assumed case – binomial sampling, in which we take a random sample of n subjects and obtain X cases. For example, to estimate the prevalence of HIV-infected subjects, we may take a random sample of ($n =$) 1000 subjects in a local community and obtain ($x =$) 5 subjects with positive results from an HIV-antibody test. In practice, however, a complete list of the sampling population needed to employ binomial sampling may not be available. We therefore discuss estimation under cluster sampling, in which the sampled unit is the cluster itself rather than the individual subject. As an example, we take a random sample of households and estimate the proportion of people who went to see a doctor in the last 12 months (Cochran, 1977). In this case, the sampled units are households rather than individuals. Other examples of the use of cluster sampling include the study of the effect of an educational intervention program on the use of solar protection among children (Mayer *et al.*, 1997) and the effect of vitamin A supplementation on child mortality (Herrera *et al.*, 1992). As noted by Cochran (1977), the estimate of the population prevalence can be subject to a large relative error when the underlying population prevalence is small under binomial sampling. Furthermore, when the disease is rare, we may even obtain 0 cases in the sample. To alleviate these concerns, we discuss the use of inverse sampling (Haldane, 1945), in which we continue sampling subjects until we obtain a predetermined number x of cases. For example, we may decide to sample subjects until we obtain, say, 5 HIV-infected cases when estimating the prevalence of HIV-infected subjects in a community. In contrast to binomial sampling, the number of cases x under inverse sampling is fixed, but the total number of sampled subjects N needed to obtain these x cases is random. Except

Statistical Estimation of Epidemiological Risk K-J. Lui
 ISBN: 0-470-85071-X (HB)

for specifically referring to the incidence rate, calculated as the number of events divided by the number of person-years of follow-up time, we will generally use the terms probability, proportion, risk, and rate synonymously in this book (Fleiss, 1981). An excellent discussion on explicit definitions of these terms as used in epidemiology appears elsewhere (Selvin, 1996).

1.1 BINOMIAL SAMPLING

Suppose that a random sample of size n is taken from a very large population so that we can reasonably assume that the probability of a randomly selected subject being a case equals a constant π and the events for each randomly selected subject of being a case or a non-case are all mutually independent. Let X denote the random number of cases among these n sampled subjects. The random variable X then follows the binomial distribution with parameters n and π:

$$\mathrm{P}(X = x|\pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x}, \tag{1.1}$$

where $x = 0, 1, \ldots, n$, $0 < \pi < 1$, and π denotes the underlying population proportion of cases. The most commonly used point estimator of the parameter π is simply the sample proportion of cases:

$$\hat{\pi} = X/n. \tag{1.2}$$

Note that under distribution (1.1), the point estimator $\hat{\pi}$ (1.2) has the expectation $\mathrm{E}(\hat{\pi}) = \pi$ (i.e., $\hat{\pi}$ is an unbiased estimator of the population proportion π) and the variance $\mathrm{Var}(\hat{\pi}) = \pi(1-\pi)/n$ (**Exercise 1.1**). In fact, the estimator $\hat{\pi}$ is the uniformly minimum variance unbiased estimator (UMVUE) of π under (1.1). By the central limit theorem, the random quantity $(\hat{\pi} - \pi)/\sqrt{\mathrm{Var}(\hat{\pi})}$ has the asymptotic standard normal distribution as $n \to \infty$. Thus, by Slutsky's theorem (Casella and Berger, 1990), we obtain an asymptotic $100(1-\alpha)$ percent confidence interval for π using Wald's statistic (Agresti and Coull, 1998),

$$[\max\{\hat{\pi} - \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}, 0\}, \quad \min\{\hat{\pi} + \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}, 1\}]. \tag{1.3}$$

Note that when $\hat{\pi} = 0$ or $\hat{\pi} = 1$, the estimated variance $\hat{\pi}(1-\hat{\pi})/n$ equals 0. Obviously, this underestimates the true variance. Therefore, whenever $\hat{\pi} = 0$ or $\hat{\pi} = 1$, we recommend use of $\hat{\pi}^*(1-\hat{\pi}^*)/n$ to estimate the variance, where $\hat{\pi}^* = (X + 0.5)/(n+1)$. Note also that although interval estimator (1.3) is easy to use, it is well known that when n is not so large that both $n\hat{\pi} \geq 5$ and $n(1-\hat{\pi}) \geq 5$ hold, (1.3) is not expected to perform well due to the possibly skewed sampling distribution of $\hat{\pi}$. To improve the performance of (1.3), we consider the probability $\mathrm{P}([(\hat{\pi} - \pi)/\sqrt{\mathrm{Var}(\hat{\pi})}]^2 \leq \mathrm{Z}^2_{\alpha/2}) \doteq 1 - \alpha$ as n is large. This leads us to obtain the

following quadratic equation (Wilson, 1927; Fleiss, 1981; Casella and Berger, 1990; Newcombe, 1998):

$$A\pi^2 - 2B\pi + C \leq 0, \tag{1.4}$$

where $A = 1 + Z^2_{\alpha/2}/n$, $B = \hat{\pi} + Z^2_{\alpha/2}/(2n)$, and $C = \hat{\pi}^2$. Because $A > 0$, (1.4) is always convex. Furthermore, we can show that $B^2 - AC > 0$ (**Exercise 1.2**) and hence the two distinct roots of $A\pi^2 - 2B\pi + C = 0$ always exist. Thus, an asymptotic $100(1-\alpha)$ percent confidence interval, which can also be derived from the score test (Wilson, 1927; Agresti and Coull, 1998; Casella and Berger, 1990; Newcombe, 1998; see also the Appendix), is given by

$$[(B - \sqrt{B^2 - AC})/A, \quad (B + \sqrt{B^2 - AC})/A]. \tag{1.5}$$

Note that an asymptotic confidence interval similar to (1.5) but with a continuity correction can be found elsewhere (Fleiss, 1981; Newcombe, 1998). Using a continuity correction can always increase the coverage probability through an increase in the length of the resulting interval estimate, but may produce a conservative confidence interval (Agresti and Coull, 1998). Note also that although interval estimator (1.5) generally outperforms (1.3), both of these confidence intervals are derived from large-sample theory. When n is small, for $X = x > 0$, we may consider using the confidence interval derived on the basis of the exact distribution (1.1) (Casella and Berger, 1990; Clopper and Pearson, 1934; Jowett, 1963):

$$\begin{aligned}[x/\{x + (n - x + 1)F_{2(n-x+1),2x,\alpha/2}\},\\ \{(x+1)F_{2(x+1),2(n-x),\alpha/2}\}/\{(n-x) + (x+1)F_{2(x+1),2(n-x),\alpha/2}\}],\end{aligned} \tag{1.6}$$

where $F_{f_1,f_2,\alpha}$ is the upper 100αth percentile of the central F distribution with f_1 and f_2 degrees of freedom. If $x = 0$, then we would define the lower limit of (1.6) to be 0. Similarly, if $x = n$, then we would define the upper limit of (1.6) to be 1. Applying interval estimator (1.6) can always guarantee the coverage probability to be larger than or equal to the desired confidence level $100(1-\alpha)$ percent for any positive integer n. Details of the derivation of confidence limits (1.6) are given in **Exercises 1.3** and **1.4**. However, it is well known that (1.6) is likely to be conservative, especially when n is not large. Blyth and Still (1983) propose another exact binomial confidence interval that satisfies a few desirable statistical properties. To facilitate the use of their interval estimator, Blyth and Still (1983) tabulate the 95% and 99% confidence limits for $n \leq 30$. They note that in some cases the interval estimate they propose can actually be contained in the resulting estimate using (1.6). Vollset (1993), Agresti and Coull (1998), and Newcombe (1998) all provide good systematic discussions comparing the performance of different interval estimators for a binomial proportion. Other closed-form interval estimators using transformations of $\hat{\pi}$ appear in **Exercises 1.5** and **1.6**.

Example 1.1 We are interested in estimating the prevalence π of subjects with hypertension in a city. Suppose that a random sample of size 200 is taken, and 35 of these 200 sampled subjects are identified to be cases. Given these data, the point estimate $\hat{\pi}$ (1.2) of the hypertension prevalence π is 0.175. The interval estimators (1.3), (1.5), and (1.6) give 95% confidence intervals for π of [0.122, 0.228], [0.129, 0.234], and [0.125, 0.235], respectively. Because both the estimates $n\hat{\pi}$ and $n(1-\hat{\pi})$ are reasonably large (at least 5), these resulting interval estimates are similar to one another; they are all appropriate for use.

Example 1.2 In a pilot study of a rare disease, suppose that we obtain only a single case with exposure to a risk factor of interest out of a random sample of 10 cases. We are interested in estimating the exposure prevalence π in the case population. Employing (1.3), (1.5), and (1.6), the corresponding 95% confidence intervals for π are [0, 0.286], [0.018, 0.404], and [0.003, 0.445]. Note that the interval estimate using (1.3) tends to shift to the left as compared with the those using (1.5) and (1.6) and therefore may not appropriate for use in this situation. It may come as no surprise that the interval estimate obtained using (1.6) is the longest of the three. This is because the coverage probability of (1.6) can be larger than the desired confidence level when n is small.

1.2 CLUSTER SAMPLING

Because of the practical difficulty of obtaining a complete list of subjects in a population, it will often be convenient to employ cluster sampling to collect data. In fact, in many circumstances clustering is unavoidable; it may even occur by study design. For example, in a study concerned with an educational intervention program on behavior change, the data are grouped into small classes (Mayer *et al.*, 1997; Lui *et al.*, 2000) and hence it is natural to treat the classes as the sampled units. When any two subjects are randomly selected from the same class, the events that these two subjects have the outcome of interest are likely to be positively correlated. Thus, the interval estimators (1.3), (1.5), and (1.6) of π, in which the intraclass correlation is not taken into account, will tend to overestimate the precision of the resulting estimate, so that the actual coverage probability of these estimators under cluster sampling will likely be less than the desired confidence level. The results presented in this section can also be useful in the situation where the measurement of the underlying response on subjects is unreliable or the cost of obtaining a new subject is much higher than obtaining a measurement from someone who is already a sampled subject (Lui, 1991). In this case, we may consider taking more than one measurement per subject to increase the efficiency or reduce the expense of a study. The number of repeated measurements taken from each subject then forms a cluster.

Suppose that a random sample of n clusters with varying cluster size $m_i (i = 1, 2, \ldots, n)$ is taken. Define $X_{ij} = 1$ if the jth $(j = 1, 2, \ldots, m_i)$ subject in the

ith cluster is a case, and $X_{ij} = 0$ otherwise. Let p_i denote the probability that a randomly selected subject from cluster i is a case; that is $\mathrm{P}(X_{ij} = 1) = p_i$ and $\mathrm{P}(X_{ij} = 0) = 1 - p_i$, where $0 < p_i < 1$. To account for the intraclass correlation between the outcomes of subjects within clusters, we assume that the p_i independently and identically follow a beta distribution beta(α, β) with mean $\pi = \alpha/T$ and variance $\pi(1-\pi)/(T+1)$, where $\mathrm{T} = \alpha + \beta$, because this family is rich in shapes and is commonly used to model Bernoulli data (Johnson and Kotz, 1969). On the basis of the above model assumptions, we can easily show that the intraclass correlation between the outcomes X_{ij} and $X_{ij'}$, $j \neq j'$, within cluster i is $\rho = 1/(T+1)$, which is always positive (**Exercise 1.7**). We can further show that the probability of a randomly selected subject being a case under the above model assumption is simply $\mathrm{E}(X_{ij}) = \mathrm{E}(\mathrm{E}(X_{ij}|p_i)) = \mathrm{E}(p_i) = \pi$.

Given p_i fixed, the conditional distribution of $X_{i\cdot} = \sum_j X_{ij}$ follows the binomial distribution with m_i and p_i. Define

$$\hat{\pi} = \sum_i X_{i\cdot}/m_{\cdot}, \tag{1.7}$$

where $m_{\cdot} = \sum_i m_i$ is the total number of sampled subjects. Note that $\hat{\pi}$ is simply the sample proportion of subjects who are the cases. We can easily show that $\hat{\pi}$ is an unbiased estimator of π under cluster sampling as well. Furthermore, we can show that the variance $\mathrm{Var}(\hat{\pi})$ (**Exercise 1.8**) is equal to

$$\mathrm{Var}(\hat{\pi}) = \pi(1-\pi)f(\mathbf{m}, \rho)/m_{\cdot}, \tag{1.8}$$

where $\mathbf{m}' = (m_1, m_2, \ldots, m_n)$ and $f(\mathbf{m}, \rho)$ is the variance inflation factor due to the intraclass correlation ρ and equals $\sum_i m_i[1 + (m_i - 1)\rho]/m_{\cdot}$, which is always greater than or equal to 1. The larger the value of ρ, the larger is the value of $f(\mathbf{m}, \rho)$. When the intraclass correlation ρ between the outcomes of all subjects within clusters equals 0, $f(\mathbf{m}, 0) = 1$ and hence the variance $\mathrm{Var}(\hat{\pi})$ reduces to $\pi(1-\pi)/m_{\cdot}$. On the other hand, when ρ equals 1, $f(\mathbf{m}, \rho)$ reaches the maximum $\sum m_i^2/m_{\cdot}$. For a given total number of subjects $m_{\cdot}$, using equal cluster size m_i will minimize the inflation factor $f(\mathbf{m}, \rho)$. To estimate ρ, we can apply the traditional intraclass correlation estimator (Fleiss, 1986; Lui *et al.*, 1996; Elston, 1977; Yamamoto and Yanagimoto, 1992)

$$\hat{\rho} = (\mathrm{BMS} - \mathrm{WMS})/[\mathrm{BMS} + (m^* - 1)\mathrm{WMS}],$$

where

$$\mathrm{BMS} = \left[\sum_i (X_{i\cdot}^2/m_i) - \left(\sum_i X_{i\cdot}\right)^2/m_{\cdot}\right] \Big/ (n-1) \text{ and}$$

$$\mathrm{WMS} = \left[\sum_i X_{i\cdot} - \sum_i (X_{i\cdot}^2/m_i)\right] \Big/ \left[\sum_i (m_i - 1)\right]$$

are the between mean-squared and within mean-squared errors, respectively, and

$$m^* = \left[\left(\sum_i m_i\right)^2 - \sum_i m_i^2\right] \Big/ \left[(n-1)\sum_i m_i\right].$$

Note that under the common correlation model (Mak, 1988), the variance formula (1.8) and the traditional intraclass correlation $\hat{\rho}$ as given above are still valid (**Exercise 1.9**).

On the basis of the above results, an asymptotic $100(1-\alpha)$ percent confidence interval for π is

$$\begin{aligned}&[\max\{\hat{\pi} - Z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})f(\mathbf{m}, \hat{\rho})/m_.}, 0\},\\ &\min\{\hat{\pi} + Z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})f(\mathbf{m}, \hat{\rho})/m_.}, 1\}].\end{aligned} \tag{1.9}$$

Note that when the cluster size $m_i = 1$ for all i, interval estimator (1.9) reduces to (1.3). Thus, when the number of subjects $m_.$ is small, (1.9), although simple to use, is unlikely to perform well. Following ideas similar to those for deriving interval estimator (1.5), we consider the following quadratic equation in π:

$$\mathcal{A}\pi^2 - 2\mathcal{B}\pi + \mathcal{C} \leq 0 \tag{1.10}$$

where $\mathcal{A} = [1 + Z_{\alpha/2}^2 f(\mathbf{m}, \hat{\rho})/m_.]$, $\mathcal{B} = [\hat{\pi} + Z_{\alpha/2}^2 f(\mathbf{m}, \hat{\rho})/(2\,m_.)]$, and $\mathcal{C} = \hat{\pi}^2$. Because $\mathcal{A} > 0$, (1.10) is always convex. Furthermore, we can show that $\mathcal{B}^2 - \mathcal{A}C > 0$ and hence an asymptotic $100(1-\alpha)$ percent confidence interval for π is given by

$$[(\mathcal{B} - \sqrt{\mathcal{B}^2 - \mathcal{A}C})/\mathcal{A},\ (\mathcal{B} + \sqrt{\mathcal{B}^2 - \mathcal{A}C})/\mathcal{A}]. \tag{1.11}$$

When $m_i = 1$ for all i, as expected, interval estimator (1.11) reduces to (1.5).

In an effort to improve the performance of (1.9), we consider use of the logarithmic transformation to improve the normal approximation of $\hat{\pi}$ (**Exercise 1.5**). By the delta method (Agresti, 1990; Casella and Berger, 1990; see also the Appendix), we can show that the asymptotic variance of $\log(\hat{\pi})$ is $\mathrm{Var}(\hat{\pi}) = (1-\pi)f(\mathbf{m}, \rho)/(m_.\pi)$. Therefore, we obtain an asymptotic $100(1-\alpha)$ percent confidence interval for π to be

$$[\hat{\pi}\exp(-Z_{\alpha/2}\sqrt{(1-\hat{\pi})f(\mathbf{m}, \hat{\rho})/(m_.\hat{\pi})}),\ \hat{\pi}\exp(Z_{\alpha/2}\sqrt{(1-\hat{\pi})f(\mathbf{m}, \hat{\rho})/(m_.\hat{\pi})})]. \tag{1.12}$$

Note that when $\hat{\pi} = 0$, $\log(\hat{\pi})$ is not defined, and when $\hat{\pi} = 1$, the estimated variance of $\log(\hat{\pi})$ is 0. In these cases, we may apply a commonly used *ad hoc* adjustment procedure for sparse data by substituting $\left(\sum X_{i.} + 0.5\right)/(m_. + 1)$ for $\hat{\pi}$ in (1.12).

Example 1.3 Consider the study of an educational intervention program on behavior change with regard to solar protection (Mayer *et al.*, 1997). There are 29 classes with sizes ranging from 1 to 6 in the intervention group and 29 classes with sizes ranging from 1 to 4 in the control groups (Lui *et al.*, 2000). Suppose that we are only interested in estimating the prevalence rate π of children who do not have an adequate level of solar protection in the intervention group. The class size and the corresponding number of children not possessing an adequate level of solar protection in this group are given in Table 1.1. The point estimate $\hat{\pi}$ in the intervention group is 0.422. Applying (1.9), (1.11), and (1.12), we obtain 95% confidence intervals for π of $[0.273, 0.570]$, $[0.286, 0.571]$, $[0.297, 0.600]$, respectively. As seen for the binomial sampling (i.e., $m_i = 1$ for all i), interval estimate (1.9) using Wald's statistic tends to shift to the left as compared with the other two estimates.

Example 1.4 A simple random sample of 30 households of size m_i ranging from 1 to 6 persons is drawn from a census taken in 1947 in wards 5 and 6 of the Eastern Health District of Baltimore (Cochran, 1977, p. 67). For each of these 30 sampled households, we ask how many persons went to see a doctor in the last 12 months. We summarize the data in Table 1.2. Suppose that we want to estimate the proportion π of people who consulted a doctor. From Table 1.2, the point estimate $\hat{\pi}$ is 0.288. The 95% confidence intervals for π obtained from (1.9), (1.11), and (1.12) are $[0.148, 0.429]$, $[0.177, 0.469]$, and $[0.172, 0.442]$, respectively. Note that if we employed the ratio estimator discussed elsewhere (Cochran, 1977) to estimate π under cluster sampling, we would obtain a 95% confidence interval for π of $[0.147, 0.430]$, which is almost the same as that obtained using (1.9), but is less preferable to interval estimates using (1.11) or (1.12).

Table 1.1 The class size and (in parentheses) the observed number of children with an inadequate level of solar protection in the intervention and control groups.

Intervention group
3(1), 2(1), 2(1), 5(0), 4(1), 3(2), 1(1), 2(2), 2(2), 2(1), 1(1), 3(2), 1(1), 3(2), 2(2), 2(0), 6(0), 2(0), 4(0), 2(1), 2(2), 2(1), 2(1), 1(1), 1(1), 1(0), 1(0), 1(0), 1(0)
Control group
2(0), 4(0), 3(2), 2(2), 3(0), 4(4), 4(2), 2(1), 2(1), 3(3), 2(2), 2(1), 4(1), 3(3), 2(2), 3(3), 1(1), 1(0), 2(1), 2(2), 2(1), 3(1), 3(2), 4(4), 1(1), 1(1), 1(1), 1(0), 1(0)

Source: Lui *et al.* (2000).

Table 1.2 Household size and (in parentheses) the observed number of people who consulted a doctor in the last 12 months for a random sample of 30 households.

5(5), 6(0), 3(2), 3(3), 2(0), 3(0), 3(0), 3(0), 4(0), 4(0), 3(0), 2(0), 7(0), 4(4), 3(1), 5(2), 4(0), 4(0), 3(1), 3(3), 4(2), 3(0), 3(0), 1(0), 2(2), 4(2), 3(0), 4(2), 2(0), 4(1)

Source: Cochran (1977).

1.3 INVERSE SAMPLING

When the underlying disease is rare (i.e., $\pi \doteq 0$), the coefficient of variation $\sqrt{(1-\pi)/(n\pi)}$ for estimator $\hat{\pi}$ (1.2) under binomial sampling (1.1) is large. Furthermore, when π is extremely small, the probability of obtaining 0 cases in our sample under (1.1) is no longer negligible for a small or even moderate sample size n. To alleviate this practical concern, we may apply inverse sampling (Haldane, 1945), in which we continue sampling subjects until we obtain a predetermined number x of cases. Let Y denote the number of non-cases before we obtain exactly x cases. The random variable Y then follows the negative binomial distribution with parameters x and π:

$$\mathrm{P}(Y = y|\pi) = \binom{x+y-1}{y} \pi^x (1-\pi)^y, \qquad y = 0, 1, 2, \ldots. \tag{1.13}$$

Under distribution (1.13), we can show that the maximum likelihood estimator (**MLE**) of π is given by

$$\hat{\pi} = x/N, \tag{1.14}$$

where $N = x + Y$. Note that (1.14) is actually a biased estimator of π. The asymptotic variance $\mathrm{Var}(\hat{\pi})$ can be shown to equal $\pi^2(1-\pi)/x$ (**Exercise 1.11**). Thus, the asymptotic coefficient of variation of $\hat{\pi}$ under distribution (1.13) is $\sqrt{(1-\pi)/x}$, which is approximately equal to $\sqrt{1/x}$ as the underlying prevalence rate $\pi \doteq 0$. In contrast to binomial sampling, we can ensure that the relative error is smaller than a given precision by simply increasing the predetermined number x of cases. Furthermore, an asymptotic $100(1-\alpha)$ percent confidence interval for π using Wald's statistic is given by

$$[\max\{\hat{\pi} - \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}^2(1-\hat{\pi})/x}, 0\},\ \min\{\hat{\pi} + \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}^2(1-\hat{\pi})/x}, 1\}]. \tag{1.15}$$

As noted before, $\hat{\pi}$ is a biased estimator of π. To alleviate this concern, for $x > 1$ we may consider use of the unbiased estimator

$$\hat{\pi}^{(\mathrm{u})} = (x-1)/(N-1), \tag{1.16}$$

which is, in fact, the UMVUE of π (**Exercise 1.12**). Best (1974) derives a closed-form expression for the variance of this estimator:

$$\mathrm{Var}(\hat{\pi}^{(\mathrm{u})}) = (x-1)(1-\pi)\left[\sum_{k=2}^{x-1}(-\pi/(1-\pi))^k/(x-k) - (-\pi/(1-\pi))^x \log(\pi)\right] - \pi^2. \tag{1.17}$$

As shown in **Exercise 1.13**, an unbiased estimator of $\mathrm{Var}(\hat{\pi}^{(\mathrm{u})})$ (1.17) for $x > 2$ is given by (Finney, 1949)

$$\widehat{\mathrm{Var}}(\hat{\pi}^{(\mathrm{u})}) = \hat{\pi}^{(\mathrm{u})}(1 - \hat{\pi}^{(\mathrm{u})})/(N-2). \tag{1.18}$$

When $x > 2$, (1.16) and (1.18) lead to an asymptotic $100(1-\alpha)$ percent confidence interval for π given by

$$\begin{aligned}&[\max\{\hat{\pi}^{(\mathrm{u})} - \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}^{(\mathrm{u})}(1-\hat{\pi}^{(\mathrm{u})})/(N-2)}, 0\},\\ &\min\{\hat{\pi}^{(\mathrm{u})} + \mathrm{Z}_{\alpha/2}\sqrt{\hat{\pi}^{(\mathrm{u})}(1-\hat{\pi}^{(\mathrm{u})})/(N-2)}, 1\}].\end{aligned} \tag{1.19}$$

Note that both interval estimators (1.15) and (1.19) may not be appropriate for use when N is not large. When N $(= x + y)$ is small, we may consider using the exact $100(1-\alpha)$ percent confidence interval $[\pi_{\mathrm{l}}^{(\mathrm{e})}, \pi_{\mathrm{u}}^{(\mathrm{e})}]$ on the basis of distribution (1.13), where $\pi_{\mathrm{l}}^{(\mathrm{e})}$ and $\pi_{\mathrm{u}}^{(\mathrm{e})}$ are the solutions of the following two equations: $\sum_{y'=0}^{y} \mathrm{P}(\mathrm{Y} = y'|\pi_{\mathrm{l}}^{(\mathrm{e})}) = \alpha/2$ and $\sum_{y'=y}^{\infty} \mathrm{P}(\mathrm{Y} = y'|\pi_{\mathrm{u}}^{(\mathrm{e})}) = \alpha/2$ (Casella and Berger, 1990). From **Exercises 1.3** and **1.14**, for $Y = y > 0$ we obtain an exact $100(1-\alpha)$ percent confidence interval for π (**Exercise 1.15**; Casella and Berger, 1990), given by

$$[x/\{x + (y+1)F_{2(y+1),2x,\alpha/2}\}, \quad xF_{2x,2y,\alpha/2}/\{xF_{2x,2y,\alpha/2} + y\}]. \tag{1.20}$$

When $y = 0$, we define the upper limit of (1.20) to be 1 for convenience. Note that the confidence limits proposed by George and Elston (1993) are actually a special case of (1.20) when $x = 1$. Lui (1995) discusses the expected length of (1.20) as a function of x and the relationship between (1.20) and the confidence limits on the expected number of trials in reliability studies previously discussed by Clemans (1959). When the underlying disease is rare (i.e., π is small), Bennett (1981) proposes an approximate $100(1-\alpha)$ percent confidence interval for π on the basis of the χ^2 distribution. Details of this can be found in **Exercise 1.20**.

Example 1.5 Suppose that we employ inverse sampling and collect 100 non-cases before obtaining exactly 20 cases. Applying interval estimators (1.15), (1.19), and (1.20), we obtain 95% confidence intervals for π of [0.100, 0.233], [0.094, 0.226], and [0.105, 0.238]. Given such an adequate number of cases, these resulting interval estimates are all similar to one another.

Example 1.6 Under inverse sampling, suppose that we decide to continue sampling subjects until we obtain exactly 2 cases. Suppose we obtain 10 non-cases in our sample. Applying interval estimators (1.15), (1.19), and (1.20), we obtain 95% confidence intervals for π of [0.000, 0.378], [0.000, 0.269], and [0.021, 0.413]. As compared with the exact 95% confidence interval, interval estimators (1.15) and (1.18), derived from large-sample theory, tend to shift to the left and are probably inadequate for use in this case.

Note that the sum $\sum_i X_i$ of independent random variables X_i, each following the binomial distribution (1.1) with parameters n_i and π, follows the binomial distribution with parameters $\sum n_i$ and π. Furthermore, the sum $\sum_i \mathrm{Y}_i$ of independent negative binomial random variables Y_i, each following the negative binomial distribution (1.13) with parameters x_i and π, follows the negative binomial with parameters $\sum_i x_i$ and π (Hoel *et al.*, 1971). Thus, in practice, we can simultaneously send several surveyors to a homogeneous population and ask each surveyor to sample a desired number of subjects n_i under binomial sampling (or continue sampling subjects until he/she has obtained a predetermined number x_i of cases under inverse sampling). We can then combine all these samples into a single database and calculate the confidence limits by simply substituting $\sum_i n_i$ for n and $\sum_i X_i$ for X under the binomial distribution (1.1) (or $\sum_i x_i$ for x and $\sum_i Y_i$ for Y under the negative binomial distribution (1.13)), respectively. All the results derived here can then be employed. Note also that when studying a rare disease in a follow-up study, we often assume that the number of cases follows a Poisson distribution. We present some useful results on estimation of the disease incidence rate under this distribution in **Exercise 1.21**. We will discuss the use of Poisson sampling in much more detail in Chapters 2 and 4.

EXERCISE

1.1. Suppose that the random variable X follows the binomial distribution (1.1) with n and π.
(a) Show that $\mathrm{E}(\hat{\pi}) = \pi$ and $\mathrm{Var}(\hat{\pi}) = \pi(1 - \pi)/n$.
(b) Find an unbiased estimator of $\mathrm{Var}(\hat{\pi})$.

1.2. Suppose that the random variable X follows the binomial distribution (1.1) with parameters n and π. Show that the inequality $B^2 - AC > 0$ always holds, where $A = 1 + \mathrm{Z}^2_{\alpha/2}/n$, $B = \hat{\pi} + \mathrm{Z}^2_{\alpha/2}/(2n)$, $C = \hat{\pi}^2$, and $\hat{\pi} = X/n$, and that the two distinct roots of $A\pi^2 - 2B\pi + C = 0$ always fall between 0 and 1 when $\hat{\pi} > 0$.

1.3. (a) Prove

$$\sum_{k=0}^{x} \binom{n}{k} \pi^k (1-\pi)^{n-k} = (n-x)\binom{n}{x} \int_0^{1-\pi} t^{n-x-1}(1-t)^x \, \mathrm{d}t.$$

(Hint: use a similar principle to mathematical induction.)
(b) Show that if F follows the F distribution with p and q degrees of freedom, then $(p/q)F/[1 + (p/q)F]$ follows the beta distribution beta$(p/2, q/2)$.
(c) On the basis of the results in (a) and (b), show that

$$\mathrm{P}(X \le x) = \mathrm{P}\left(F > \frac{(n-x)\pi}{(x+1)(1-\pi)}\right),$$

where X is binomial with parameters n and π, and $F \sim F_{2(x+1),2(n-x)}$, respectively.

1.4. Based on the result in part (c) of **Exercise 1.3**, derive the confidence limits (1.6).

1.5. Using the delta method (Agresti, 1990), show that the asymptotic variance of $\log(\hat{\pi})$ is $(1-\pi)/(n\pi)$ under distribution (1.1) and discuss how to apply this result to derive an asymptotic $100(1-\alpha)$ percent confidence interval for π.

1.6. Using the delta method, show that the asymptotic variance of $2\sin^{-1}\sqrt{\hat{\pi}}$ is $1/n$ under distribution (1.1) and discuss how to apply this result to derive an asymptotic $100(1-\alpha)$ percent confidence interval for π.

1.7. Suppose that the Bernoulli random variable X_{ij} has the probability mass function $P(X_{ij}=1)=p_i$ and $P(X_{ij}=0)=1-p_i$, where p_i follows the beta distribution with mean $E(p_i)=\pi$ and variance $\pi(1-\pi)/(T+1)(i=1,2,..,n, j=1,2,\ldots,m_i)$. Suppose further that, given p_i fixed, X_{ij} and $X_{ij'}$ are conditionally independent for $j\neq j'$. Show that the intraclass correlation between X_{ij} and $X_{ij'}$ (where $j\neq j'$) within cluster i is $\rho=1/(T+1)$.

1.8. Show that the variance of $\hat{\pi}$ $(=\sum_{i=1}^{n} X_{i\cdot}/m_{\cdot}$, where $X_{i\cdot}=\sum_{j=1}^{m_i} X_{ij}$ and $m_{\cdot}=\sum_i m_i)$ under the model assumption in **Exercise 1.7** is $\mathrm{Var}(\hat{\pi})=\pi(1-\pi)f(\mathbf{m},\rho)/m_{\cdot}$, where $\mathbf{m}'=(m_1,m_2,\ldots,m_n)$ and $f(\mathbf{m},\rho)=\sum_i m_i[1+(m_i-1)\rho]/m_{\cdot}$.

1.9. Under the common correlation model, we assume that the joint probabilities of any two different dichotomous responses X_{ij} and $X_{ij'}$ within a given cluster i are defined as follows:

$$P(X_{ij}=1, X_{ij'}=1)=\pi^2+\rho\pi(1-\pi),$$
$$P(X_{ij}=0, X_{ij'}=0)=(1-\pi)[(1-\pi)+\rho\pi],$$
$$P(X_{ij}=0, X_{ij'}=1)=P(X_{ij}=1, X_{ij'}=0)=\pi(1-\pi)(1-\rho).$$

(a) Show that the intraclass correlation between X_{ij} and $X_{ij'}$ is equal to ρ.
(b) Show that the variance $\mathrm{Var}(\hat{\pi})$ (where $\hat{\pi}=\sum_{i=1}^{n} X_{i\cdot}/m_{\cdot}$) is equal to $\pi(1-\pi)f(\mathbf{m},\rho)/m_{\cdot}$, where $f(\mathbf{m},\rho)=\sum_i m_i[1+(m_i-1)\rho]/m_{\cdot}$. This is actually the same as that given under the beta-binomial model for $\rho=1/(T+1)$.
(c) Show that the expectation $E(\mathrm{WMS})=\pi(1-\pi)(1-\rho)$ and $E(\mathrm{BMS})=\pi(1-\pi)(1-\rho)+m^*\pi(1-\pi)\rho$, where $m^*=\left[\left(\sum_i m_i\right)^2-\sum_i m_i^2\right]\Big/\left[(n-1)\sum_i m_i\right]$. Thus, we may apply the traditional intraclass correlation estimator $\hat{\rho}=(\mathrm{BMS}-\mathrm{WMS})/[\mathrm{BMS}+(m^*-1)\mathrm{WMS}]$ to estimate ρ.

1.10. Consider the data for the control group in Table 1.1. (a) What is the MLE $\hat{\pi}$ of the prevalence of children with an inadequate level of solar protection in the control group? (b) What are the corresponding 95% confidence intervals for π using (1.9), (1.11), and (1.12)?

1.11. Show that the asymptotic variance $\mathrm{Var}(\hat{\pi})$ under distribution (1.13) is $\pi^2(1-\pi)/x$, where $\hat{\pi}$ is given in (1.14). Thus, the coefficient of variation of $\hat{\pi}$ is $\sqrt{(1-\pi)/x}$.

1.12. Show that $\hat{\pi}^{(\mathrm{u})}(=(x-1)/(N-1))$ is an unbiased estimator of π under distribution (1.13). Note that because $\hat{\pi}^{(\mathrm{u})}$ is a function of complete sufficient statistic, $\hat{\pi}^{(\mathrm{u})}$ is the UMVUE of π (Casella and Berger, 1990).

1.13. Show that for $x > 2$, we have $\mathrm{E}[\hat{\pi}^{(\mathrm{u})}(1-\hat{\pi}^{(\mathrm{u})})/(N-2)] = \mathrm{Var}(\hat{\pi}^{(\mathrm{u})})$, where the expectation is taken with respect to distribution (1.13) (Finney, 1949).

1.14. Show that the cumulative distribution $\sum_{y'=0}^{y} \mathrm{P}(Y = y'|\pi)$, where Y follows the negative binomial distribution (1.13) with parameters x and π, equals $\sum_{x'=0}^{y} \mathrm{P}(X = x'|1-\pi)$, where X has the binomial distribution with parameters $n = x + y$ and $1 - \pi$ (Morris, 1963).

1.15. On the basis of the results found in **Exercises 1.3** and **1.14**, for $y > 0$ derive the $100(1-\alpha)$ percent confidence limits $[\pi_{\mathrm{l}}^{(\mathrm{e})}, \pi_{\mathrm{u}}^{(\mathrm{e})}]$ given in (1.20) where $\pi_{\mathrm{l}}^{(\mathrm{e})} = x/[x + (y+1)F_{2(y+1),2x,\alpha/2}]$, and $\pi_{\mathrm{u}}^{(\mathrm{e})} = xF_{2x,2y,\alpha/2}/[xF_{2x,2y,\alpha/2} + y]$, respectively.

1.16. Consider an experiment consisting of x randomly selected devices (where x is fixed), each subject to a series of independent and identical trials until it fails. Suppose that the failure probability at each trial equals a constant π and all these failures between trials are mutually independent. Discuss how we can apply formula (1.20) to derive a $100(1-\alpha)$ percent confidence interval for the expected number of trials $(1-\pi)/\pi$ before the failure of a given device. (Hint: $f(\pi) = (1-\pi)/\pi$ is a monotonically decreasing function of π.)

1.17. Using the delta method, show that $2\sinh^{-1}(\sqrt{Y/x})$, where Y follows the negative binomial distribution (1.13) has the asymptotic variance $1/x$. Thus, the transformation $2\sinh^{-1}(\sqrt{Y/x})$ can be used to stabilize the variance of the negative binomial random variable Y.

1.18. Suppose that a random sample of size 1000 subjects is taken. Suppose further that we find 5 cases in this sample. What are the 95% confidence intervals for the prevalence of cases using (1.3), (1.5), and (1.6)?

1.19. Let Y_i denote the number of trials before failure for device $i (i = 1, 2, \ldots, 5)$ in **Exercise 1.16**. Suppose that $\sum Y_i = 100$. What is the 95% confidence interval for the expected number of trials $(1-\pi)/\pi$ before failure for a given device.

1.20. When the underlying prevalence rate π is small, show that $2(x+Y)\pi$, where Y follows the negative binomial distribution (1.13), follows approximately the χ^2 distribution with $2x$ degrees of freedom (Bennett, 1981). How can we apply this result to derive an approximate $100(1-\alpha)$ percent confidence interval for π?

1.21. When the underlying disease incidence rate λ is small in cohort studies, the number X of cases is often assumed to follow a Poisson distribution: $\exp(-\lambda n^*)(\lambda n^*)^X/X!$, where n^* is a known total of follow-up time in person-years and $X = 0, 1, 2, \ldots$.
(a) Show that $\hat{\lambda} = X/n^*$ is the MLE and an unbiased estimator of λ with variance λ/n^*.
(b) Show that an asymptotic $100(1-\alpha)$ percent confidence interval for λ is given by $\hat{\lambda} \pm Z_{\alpha/2}\sqrt{\hat{\lambda}/n^*}$.
(c) Show that an asymptotic $100(1-\alpha)$ percent confidence interval for λ can be given by solving the two distinct roots of the following quadratic equation: $A^\dagger\lambda^2 - 2B^\dagger\lambda + C^\dagger = 0$, $A^\dagger = 1$, $B^\dagger = \hat{\lambda} + Z^2_{\alpha/2}/(2n^*)$, and $C^\dagger = \hat{\lambda}^2$.
(d) Using the fact that

$$\sum_{X=0}^{x_0} \exp(-\lambda n^*)(\lambda n^*)^X/X! = \mathrm{P}(\chi^2_{2(x_0+1)} > 2n^*\lambda),$$

where $\chi^2_{2(x_0+1)}$ is a chi-squared random variable with $2(x_0+1)$ degrees of freedom (Casella and Berger, 1990), show that an exact $100(1-\alpha)$ percent confidence interval for λ is $[\chi^2_{2x_0,1-\alpha/2}/(2n^*), \chi^2_{2(x_0+1),\alpha/2}/(2n^*)]$, where $\chi^2_{f,\alpha}$ is the upper 100αth percentile of the central χ^2 distribution with f degrees of freedom. Note that if $x_0 = 0$, we define the lower limit to be 0.)

1.22. Suppose that in **Exercise 1.21**, we follow a group of 25 subjects for 2 years and obtain $X = 10$ cases.
(a) What is the MLE $\hat{\lambda}$ of the disease incidence rate λ?
(b) What is the 95% confidence interval for λ using Wald's statistic?
(c) What is the 95% confidence interval for λ using the quadratic equation in **Exercise 1.21**?
(d) What is the exact 95% confidence interval for λ?

REFERENCES

Agresti, A. (1990) *Categorical Data Analysis*. Wiley, New York.

Agresti, A. and Coull, B. A. (1998) Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician*, **52**, 119–126.

Bennett, B. M. (1981) On the use of the negative binomial in epidemiology. *Biometrical Journal*, **23**, 69–72.

Best, D. J. (1974) The variance of the inverse binomial estimator. *Biometrika*, **67**, 385–386.

Blyth, C. R. and Still, H. A. (1983) Binomial confidence intervals. *Journal of the American Statistical Association*, **78**, 108–116.

Casella, G. and Berger, R. L. (1990) *Statistical Inference*. Duxbury, Belmont, CA.

Clemans, K. G. (1959) Confidence limits in the case of the geometric distribution. *Biometrika*, **46**, 260–264.

Clopper, C. J. and Pearson, E. S. (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, **26**, 404–413.

Cochran, W. G. (1977) *Sampling Techniques*, 3rd edition. Wiley, New York.

Elston, R. C. (1977) Response to query: Estimating 'heritability' of a dichotomous trait. *Biometrics*, **33**, 232–233.
Finney, D. J. (1949) On a method of estimating frequencies. *Biometrika*, **36**, 233–234.
Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. Wiley, New York.
Fleiss, J. L. (1986) *The Design and Analysis of Clinical Experiments*, Wiley, New York.
George, V. T. and Elston R. C. (1993) Confidence limits based on the first occurrence of an event. *Statistics in Medicine*, **12**, 685–690.
Haldane, J. B. S. (1945) On a method of estimating frequencies. *Biometrika*, **33**, 222–225.
Herrera, M. G., Nestel, P., Amin, A. E., Fawzi, W. W., Mohamed, K. A. and Weld, L. (1992) Vitamin A supplementation and child survival. *Lancet*, **340**, 267–271.
Hoel, P. G., Port, S. C. and Stone, C. J. (1971) *Introduction to Probability Theory*. Houghton Mifflin, Boston.
Johnson, N. L. and Kotz, S. (1969) *Distributions in Statistics: Discrete Distributions*. Wiley, New York.
Jowett, G. H. (1963) The relationship between the binomial and *F* distributions. *The Statistician*, **13**, 55–57.
Lui, K. -J. (1991) Sample size for repeated measurements in dichotomous data. *Statistics in Medicine*, **10**, 463–472.
Lui, K. -J. (1995) Confidence limits for the population prevalence rate based on the negative binomial distribution. *Statistics in Medicine*, **14**, 1471–1477.
Lui, K. -J., Cumberland, W. G. and Kuo, L. (1996) An interval estimate for the intraclass correlation in beta-binomial sampling. *Biometrics*, **52**, 412–425.
Lui, K. -J., Mayer, J. A. and Eckhardt, L. (2000) Confidence intervals for the risk ratio under cluster sampling based on the beta-binomial model. *Statistics in Medicine*, **19**, 2933–2942.
Mak, T. K. (1988) Analysing intraclass correlation for dichotomous variables. *Applied Statistics*, **37**, 344–352.
Mayer, J. A., Slymen, D. J., Eckhardt, L., *et al.* (1997) Reducing ultraviolet radiation exposure in children. *Preventive Medicine*, **26**, 516–522.
Morris, K. W. (1963) A note on direct and inverse binomial sampling. *Biometrika*, **50**, 544–545.
Newcombe, R. G. (1998) Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, **17**, 857–872.
Selvin, S. (1996) *Statistical Analysis of Epidemiologic Data*. Oxford University Press, New York.
Vollset, S. E. (1993) Confidence intervals for a binomial proportion. *Statistics in Medicine*, **12**, 809–824.
Wilson, E. B. (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.
Yamamoto, E. and Yanagimoto, T. (1992) Moment estimators for the beta-binomial distribution. *Journal of Applied Statistics*, **19**, 273–283.