Journal of Official Statistics, Vol. 16, No. 1, 2000, pp. 31–37

A Note on the Use of Inverse Sampling: Point Estimation Between Successive Infections

Kung-Jong Lui[1]

This article considers point estimation of the effect of the primary infection on the likelihood of developing a secondary infection. The article notes that the use of inverse sampling can avoid the theoretical limitations, such as bias or nonexistence of variance, in application of the maximum likelihood estimator under the commonly-assumed multinomial sampling. Under inverse sampling, this article derives the uniformly minimum variance unbiased estimators (UMVUEs) of the risk ratio and risk difference, as well as their corresponding variances. This article further develops the UMVUEs of these variances as well. Finally, the article includes a discussion on interval estimation and applies Monte Carlo simulation to evaluate and compare the performance of the interval estimators that are derived from the point and variance estimators obtained here.

Key words: Point estimation; uniformly minimum variance unbiased estimator; maximum likelihood estimator; inverse sampling; risk ratio; risk difference.

1. Introduction

Consider a study assessing the effect of the primary infection on the likelihood of developing a secondary infection (Agresti 1990). We may wish to estimate the risk ratio (RR) or risk difference (RD) (Fleiss 1981) between the secondary infection, given the primary infection, and the primary infection. Under the multinomial sampling, in which the total number of subjects in the studies is fixed (Lui 1998), one can easily show that the maximum likelihood estimators (MLEs) of these two important epidemiologic indices have infinitely large bias and no exact variances. To avoid these theoretical limitations, this article proposes the use of inverse sampling (Haldane 1945) and notes that the uniformly minimum variance unbiased estimators (UMVUEs) of RR and RD, as well as their exact variances can be easily derived. This article further develops the UMVUEs of these variances as well. Other discussions on point estimation under inverse sampling in situations different from those treated here appear elsewhere (Bennett 1981, 1986; Roberts 1988; Singh and Aggarwal 1991; Lui 1996a, 1996b, 1997a). Discussion of interval estimation (rather than point estimation) of the RR between successive infections under multinomial sampling (rather than inverse sampling) can be found elsewhere as well (Lui 1998).

[1] Department of Mathematical and Computer Sciences, San Diego State University, San Diego, CA 92182-7720, U.S.A. E-mail: Kjl@rohan.sdsu.edu

Acknowledgments: The author wishes to thank the Editor and the two referees for valuable comments on improving the clarity of this article.

2. Theory and Estimators

Consider a study of a disease epidemic in which randomly selected subjects from a population are first classified according to whether they had contracted a primary infection and then reclassified according to whether they had developed a secondary infection within a given time period after clearing up the first infection. For clarity, I use the following four-fold table to summarize the data structure:

		Secondary Infection: Yes	Secondary Infection: No	
Primary Infection	Yes	p_{11}	p_{12}	$p_{1.}$
	No	–	p_{22}	p_{22}

where p_{11}, p_{12}, and p_{22} denote the cell probabilities, $0 < p_{ij} < 1$, $p_{1.} = p_{11} + p_{12}$, and $p_{22} = 1 - p_{1.}$. Note that, by definition, a subject cannot have a secondary infection without first having a primary infection and hence no subjects can fall in the cell with no primary infection but with a secondary infection.

To assess the magnitude of the effect of the primary infection on the likelihood of developing the secondary infection, we wish to estimate the *RR*, defined as $(p_{11}/p_{1.})/p_{1.}$, or the *RD*, defined as $(p_{11}/p_{1.}) - p_{1.}$, between the secondary infection, given the primary infection, and the primary infection. Note that under the commonly-assumed multinomial sampling scheme, in which the total number of subjects n in a study is fixed, the probability mass function for the random vector (n_{11}, n_{12}, n_{22}) is $f(n_{11}, n_{12}, n_{22}) = n!/(n_{11}!n_{12}!n_{22}!)\ p_{11}^{n_{11}}\ p_{12}^{n_{12}}\ p_{22}^{n_{22}}$, where n_{ij} is the observed frequency corresponding to the cell with probability p_{ij}, $0 < p_{ij} < 1$, $p_{11} + p_{12} + p_{22} = 1$, and $n_{11} + n_{12} + n_{22} = n$. Because both p_{11} and p_{12} are > 0, the probability $p_{1.}$ is always > 0. Thus, the above *RR* and *RD* are always well-defined. The MLEs of the *RR* and *RD* under the multinomial sampling are simply $(\hat{p}_{11}/\hat{p}_{1.})/\hat{p}_{1.}$ and $(\hat{p}_{11}/\hat{p}_{1.}) - \hat{p}_{1.}$, respectively, where $\hat{p}_{11} = n_{11}/n$, $\hat{p}_{1.} = n_{1.}/n, n_{1.} = n_{11} + n_{12}$. Note that because $n_{1.}$ (or equivalently, $\hat{p}_{1.}$) is a random variable under this sampling scheme and can be 0 with a positive probability, the biases of these MLEs are ∞. Furthermore their variances do not exist. In practice, if $n_{1.}$ should equal 0 under the multinomial scheme, we may commonly apply the adjustment procedure for sparse data by adding 0.50 to each cell when calculating $(\hat{p}_{11}/\hat{p}_{1.})/\hat{p}_{1.}$ or $(\hat{p}_{11}/\hat{p}_{1.}) - \hat{p}_{1.}$. However, this adjustment procedure is somewhat ad hoc and its optimal statistical properties are difficult to justify. Furthermore, the bias and variance of the resulting estimators are also difficult to derive explicitly, although both are finite when using the above adjustment procedure.

To avoid the undesirable situations in which the MLEs of the *RR* and *RD* are not defined under the multinomial sampling scheme, we may consider fixing $n_{1.}$ to assure that $n_{1.} > 0$. This leads us to consider use of inverse sampling (Haldane 1945), in which we continue sampling subjects until we obtain a pre-determined fixed number $n_{1.}$ of subjects with a primary infection. Because it may take time to develop the underlying disease of interest, to avoid the practical difficulty of employing inverse sampling, we may want to apply this sampling scheme at the end of the first attack period. Let n_{11} denote the number of subjects

later developing infection among these $n_{1.}$ subjects with the primary infection and let n_{22} denote the number of subjects without the primary infection needed to collect the desired number of $n_{1.}$ subjects with the primary infection. The joint probability mass function for (n_{11}, n_{22}) is then given by

$$\begin{aligned} f(n_{11}, n_{22}) &= \binom{n_{1.}}{n_{11}} \left(\frac{p_{11}}{p_{1.}}\right)^{n_{11}} \left(1 - \frac{p_{11}}{p_{1.}}\right)^{n_{1.}-n_{11}} \binom{n_{22}+n_{1.}-1}{n_{22}} {p_{1.}}^{n_{1.}} (1-p_{1.})^{n_{22}} \\ &= \binom{n_{1.}}{n_{11}} \binom{n_{22}+n_{1,}-1}{n_{22}} {p_{11}}^{n_{11}} {p_{12}}^{n_{1.}-n_{11}} {p_{22}}^{n_{22}} \end{aligned} \tag{1}$$

where $n_{11} = 0, 1, 2, \ldots, n_{1.}$, and $n_{22} = 0, 1, 2, \ldots$.

On the basis of (1), note that n_{11} and n_{22} are independent. Note further that by use of Theorem 5.6 on page 46 in the textbook by Lehmann (1983), (n_{11}, n_{22}) is, in fact, a complete sufficient statistic.

Define

$$\hat{RR} = \left(\frac{n_{11}}{n_{1.}}\right)\left(\frac{n_{22}+n_{1.}}{n_{1.}}\right) \tag{2}$$

Because $E(\hat{RR}) = E\big(n_{11}/n_{1.}\big)E\big((n_{22}+n_{1.})/n_{1.}\big) = (p_{11}/p_{1.})/p_{1.}$, $\hat{RR}$ (2) is the UMVUE of RR (Lehmann 1983; Casella and Berger 1990). Furthermore, I can show that the variance $\mathrm{Var}(\hat{RR})$ is

$$\mathrm{Var}(\hat{RR}) = p_{11}p_{12}(1-p_{1.})/(n_{1.}^2 p_{1.}^4) + p_{11}^2(1-p_{1.})/(n_{1.}p_{1.}^4) + p_{11}p_{12}/(n_{1.}p_{1.}^4). \tag{3}$$

I also derive the UMVUE $\mathrm{Var}(\hat{RR})$ of this variance (3) and present the result in the Appendix.

Similarly, the UMVUE of *RD* is given by

$$\hat{RD} = \left(\frac{n_{11}}{n_{1.}}\right) - \hat{p}_{1.}^{*} \tag{4}$$

where $\hat{p}_{1.}^{*} = (n_{1.}-1)/(n_{1.}+n_{22}-1)$ is an unbiased estimator $p_{1.}$ under inverse sampling when $n_{1.} > 2$. Note that Best (1974) shows that

$$\begin{aligned} \mathrm{Var}(\hat{p}_{1.}^{*}) &= (n_{1.}-1)(1-p_{1.}) \\ &\quad \times \left\{ \sum_{k=2}^{n_{1.}-1} (-p_{1.}/(1-p_{1.}))^k/(n_{1.}-k) - (-p_{1.}/(1-p_{1.}))^{n_{1.}} \log(p_{1.}) \right\} - p_{1.}^2 \end{aligned}$$

Hence, the variance of $\hat{RD}$ is

$$\mathrm{Var}(\hat{RD}) = p_{11}p_{12}/(n_{1.}p_{1.}^2) + \mathrm{Var}(\hat{p}_{1.}^{*}) \tag{5}$$

Furthermore, I derive and present the UMVUE $\hat{\mathrm{Var}}(\hat{RD})$ of variance (5) in the Appendix as well.

3. An Example

Consider the example of a sample of calves in Florida (Agresti 1990). For illustration purpose only, assume that we obtain the same data as shown on page 46 of Agresti (1990) by employing inverse sampling, in which we collect $n_{22} = 63$ calves with no pneumonia infection before we obtain the first $n_{1.} = 93$ calves (which is pre-determined and fixed)

with a primary pneumonia infection. We further assume that among these 93 calves, $n_{11} = 30$ calves later develop the secondary infection within two weeks after the first infection clears up. On the basis of these data, the UMVUE $\hat{RR}$ in (2) is 0.541 with the resulting variance estimate $\hat{\text{Var}}(\hat{RR})$ in (A.1) equal to 0.0079. Similarly, when applying the UMVUEs $\hat{RD}$ in (4) and $\hat{\text{Var}}(\hat{RD})$ in (A.2) to estimate the RD and $\text{Var}(\hat{RD})$, we obtain -0.2710 and 0.0039, respectively. Both results suggest that the primary infection generates a natural immunity to reduce the likelihood of developing the secondary pneumonia infection.

4. Discussion

To study the effect of the primary infection on the likelihood of developing the secondary infection, it is certainly not ethical to employ an experimental design, in which one can randomly assign subjects to one of the two comparison groups. Instead, we need to rely on the cohort study design (Fleiss 1981). This article provides a sampling design to study natural immunity, while avoiding the theoretical difficulty in point estimation of the RR and the RD under multinomial sampling design.

Note that the estimator $\hat{RR}$ in (2) is actually the MLEs under both the multinomial and the inverse samplings considered here. However, while this estimator is biased with no exact variance under the former sampling, use of the latter sampling can easily avoid these theoretical limitations. Furthermore, the MLE of RD under both of these sampling schemes is $(n_{11}/n_{1.}) - \hat{p}_{1.}$, where $\hat{p}_{1.} = n_{1.}/(n_{1.} + n_{22})$, which is slightly different from $\hat{RD}$ (4) and is a biased estimator of the RD under the sampling scheme proposed here.

The property of unbiasedness is not invariant through the reciprocal transformation. The inverse of the UMVUE $\hat{RR}$ is certainly not the UMVUE of $1/RR$. On the other hand, this concern does not apply to the case of point estimation for the RD; i.e., $-\hat{RD}$ is obviously the UMVUE of $-RD = p_{1.} - (p_{11}/p_{1.})$.

Finally, note that the aim of point estimation is to find an estimator with a small bias and a small variance. When we need to give an estimate of a parameter as a single value, an interval estimator that provides a set of values is certainly not appropriate to meet this need. On the other hand, we may also wish to find out if we can derive good interval estimators from the point estimators obtained here. Instead of discussing the bias and the variance for point estimation, we commonly consider the coverage probability and the average length to evaluate the performance of an interval estimator (Casella and Berger 1990).

Suppose that we want to provide a $100(1 - \alpha)\%$ confidence interval of RR. The easiest and most naive method on the basis of (2) is given by

$$\hat{RR} + Z_{\alpha/2}\sqrt{\hat{\text{Var}}(\hat{RR})} \tag{6}$$

where $\hat{\text{Var}}(\hat{RR})$ is given in (A.1) in the Appendix and Z_α is the upper (100α)th percentile of the standard normal distribution. However, as noted elsewhere in other situations (Jewell 1986; Lui 1996a), since the sampling distribution of the UMVUE $\hat{RR}$ (2) can be skewed, an interval estimator (6) does not necessarily perform well, especially when the expected number of subjects is not reasonably large in all cells. Thus, to improve the normal approximation of the sampling distribution of $\hat{RR}$, we may consider use of the logarithmic transformation (Katz, Baptista, Azen, and Pike 1978; Lui 1998). By using the delta method (Casella and Berger 1990), we obtain the estimated asymptotic variance of $\log(\hat{RR})$ which is

$\hat{\text{V}}\text{ar}(\hat{RR})/(\hat{\text{RR}})^2$. Hence, a $100(1-\alpha)\%$ confidence interval of the *RR* is given by

$$(RR_l, RR_u) \tag{7}$$

where $RR_l = \exp(\log(\hat{RR}) - Z_{\alpha/2}\sqrt{\hat{\text{V}}\text{ar}(\hat{RR})/(\hat{RR})^2})$ and $RR_u = \exp(\log(\hat{RR}) + Z_{\alpha/2}\sqrt{\hat{\text{V}}\text{ar}(\hat{RR})/(\hat{\text{RR}})^2})$, respectively. Note that when $\hat{RR} = 0, \log(\hat{RR})$ is not defined. To eliminate this difficulty, we may apply the commonly-used ad hoc adjustment procedure for sparse data by adding 0.50 to each cell whenever $n_{11} = 0$. To compare and evaluate the finite sample performance of these two interval estimators (6 and 7), I use Monte Carlo simulation. I apply SAS (1990) to generate the desired random observations according to the distribution defined in (1). Note that for given *RR* and $p_{1.}$, the cell probabilities p_{ij} are all uniquely determined: $p_{11} = RRp_{1.}^2$, $p_{12} = p_{1.} - p_{11}$, and $p_{22} = 1 - p_{1.}$. I consider the situations in which the probability of the primary infection $p_{1.} = 0.2, 0.3, 0.5$; the underlying risk ratio $RR = 0.25, 0.50$, and 1; the number of subjects with the primary infection $n_{1.} = 20, 30$, and 50. For each configuration determined by a combination of these parameters, I generated 10,000 repeated samples to estimate the coverage probability and the average length of the 95% confidence interval. On the basis of the simulated results (which are not presented here for brevity, but are available to readers upon request), I have found that when the expected number of subjects with the secondary infection ($= n_{1.}RRp_{1.}^2$) is small, the coverage probability of interval estimator (6) can be much less than the desired confidence level. I also have found that interval estimator (7) using the logarithmic transformation not only outperforms estimator (6), but also consistently performs well (i.e., the estimated coverage probabilities range from 95% to 97%) in all the situations considered here. Thus, I recommend interval estimator (7) for general use. Because the performance of an interval estimator depends heavily on the sampling distribution of the statistic we employ, as shown here, we should not directly apply the UMVUE and its estimated variance as in interval estimator (6) to produce an interval estimate without first attempting to improve the normal approximation through an appropriate transformation.

In summary, this article notes the usefulness of inverse sampling when we do point estimation of the *RR* and the *RD* between successive infections. This article derives the UMVUEs of these two important epidemiological indices and their respective variances in closed form. This article further derives the UMVUEs of these variances. Finally, this article includes a discussion on interval estimation and applies Monte Carlo simulation to evaluate the performance on interval estimators that are derived from the point and variance estimators obtained here.

Appendix

Following Finney (1949) and Lehmann (1983), we can show the following expectations with respect to distribution (1):

$$E[n_{22}(n_{1.} + n_{22})/\{n_{1.}(n_{1.} + 1)\}] = (1 - p_{1.})/p_{1.}^2$$

$$E[(n_{1.} + n_{22})(n_{1.} + n_{22} + 1)/\{n_{1.}(n_{1.} + 1)\}] = 1/p_{1.}^2 \text{ and}$$

$$E\left[n_{1.}\left\{\left(\frac{n_{11}}{n_{1.}}\right) - \left(\frac{n_{11}}{n_{1.}}\right)^2\right\}/(n_{1.} - 1)\right] = p_{11}p_{12}/p_{1.}^2$$

On the basis of these results and the independence between n_{11} and n_{22}, the UMVUE of $\mathrm{Var}(\hat{RR})$ is $\hat{\mathrm{Var}}(\hat{RR}) =$

$$\frac{1}{n_{1.}^2}\left[n_{1.}\left\{\left(\frac{n_{11}}{n_{1.}}\right)-\left(\frac{n_{11}}{n_{1.}}\right)^2\right\}\Big/(n_{1.}-1)\right][n_{22}(n_{1.}+n_{22})/\{n_{1.}(n_{1.}+1)\}]$$
$$+\frac{1}{n_{1.}}\left[\left(\frac{n_{11}}{n_{1.}}\right)^2-\left\{\left(\frac{n_{11}}{n_{1.}}\right)-\left(\frac{n_{11}}{n_{1.}}\right)^2\right\}\Big/(n_{1.}-1)\right][n_{22}(n_{1.}+n_{22})/\{n_{1.}(n_{1.}+1)\}]$$
$$+\frac{1}{n_{1.}}\left[n_{1.}\left\{\left(\frac{n_{11}}{n_{1.}}\right)-\left(\frac{n_{11}}{n_{1.}}\right)^2\right\}\Big/(n_{1.}-1)\right][(n_{1.}+n_{22})(n_{1.}+n_{22}+1)/\{n_{1.}(n_{1.}+1)\}] \tag{A1}$$

Furthermore, Finney (1940) shows that $E\{\hat{p}_{1.}^*(1-\hat{p}_{1.}^*)/(n_{1.}+n_{22}-2)\} = \mathrm{Var}(\hat{p}_{1.}^*)$ for $n_{1.} > 2$, where $\hat{p}_{1.}^* = (n_{1.}-1)/(n_{1.}+n_{22}-1)$. Thus, the UMVUE of $\mathrm{Var}(\hat{RD})$ is simply

$$\hat{\mathrm{Var}}(\hat{RD}) = \hat{p}_{1.}^*(1-\hat{p}_{1.}^*)/(n_{1.}+n_{22}-2)+\left\{\left(\frac{n_{11}}{n_{1.}}\right)-\left(\frac{n_{11}}{n_{1.}}\right)^2\right\}/(n_{1.}-1) \tag{A2}$$

5. References

Agresti, A. (1990). Categorical Data Analysis. New York: Wiley & Sons.

Bennett, B.M. (1981). On the Use of the Negative Binomial in Epidemiology. Biometrical Journal, 23, 69–72.

Bennett, B.M. (1986). On Combining Estimates of Relative Risk Using the Negative Binomial Model. Biometrical Journal, 28, 859–862.

Best, D.J. (1974). The Variance of the Inverse Binomial Estimator. Biometrika, 61, 385–386.

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

Finney, D.J. (1949). On a Method of Estimating Frequencies. Biometrika, 36, 233–234.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, 2nd edn. New York: Wiley & Sons.

Haldane, J.B.S. (1945). On a Method of Estimating Frequencies. Biometrika, 33, 222–225.

Jewell, N.P. (1986). On the Bias of Commonly Used Measures of Association for 2×2 Tables. Biometrics, 42, 351–358.

Katz, D., Baptista, J., Azen, S.P., and Pike, M.C. (1978). Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. Biometrics, 34, 469–474.

Lehmann, E.L. (1983). Theory of Point Estimation. New York: Wiley & Sons.

Lui, K.-J. (1996a). Point Estimation on Relative Risk under Inverse Sampling. Biometrical Journal, 38, 669–680.

Lui, K.-J. (1996b). Notes in Case-Control Studies with Matched-Pairs under Inverse Sampling. Biometrical Journal, 38, 681–693.

Lui, K.-J. (1997a). Conditional Estimation and Exact Test of the Common Relative Difference in Combination of 2×2 Tables under Inverse Sampling. Biometrical Journal, 39, 215–225.

Lui, K.-J. (1998). Interval Estimation of the Risk Ratio between the Secondary Infection Given the Primary Infection and the Primary Infection. Biometrics, 54, 706–711.

Roberts, C. (1988). Unbiased Estimation of Relative Risk Using a Binomial with a Negative Binomial Plan. Biometrical Journal, 30, 939–944.

SAS (1990). SAS Language, Reference Version 6, 1st edition. Cary, North Carolina: SAS Institute, Inc.

Singh, P. and Aggarwal, A.R. (1991). Inverse Sampling in Case Control Studies. Environmetrics, 2, 293–299.

Received January 1998
Revised February 1999