Mathematical Statistics
Stockholm University

Compensating for incomplete ascertainment when calculating cluster measures

Åke Svensson

Research Report 2002:6

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.matematik.su.se/matstat

Compensating for incomplete ascertainment when calculating cluster measures

Åke Svensson

University of Stockholm and Swedish Institute for Infectious Disease Control

August 13, 2002

Abstract

In many applications individuals are divided into groups or clusters depending on individual characteristics. Often cluster measures based on the cluster size distribution are calculated. Incomplete ascertainment of individuals makes the observed cluster xsizes smaller than the actual sizes. A method to derive estimates that compensates for possible bias, when the ascertainment is random and the ascertainment probability is known, is suggested. The method is applied to problems in tuberculosis epidemiology. In some examples it is shown that the randomness caused by the ascertainment is larger than the bias.

1. Introduction

This paper is concerned with the problem of incomplete ascertainment of individuals that are divided into separate groups or clusters. In many cases the distribution of the clusters sizes is an interesting subject for analysis. If not all members of a cluster are identified the observed cluster sizes will always be smaller than the actual sizes. The present discussion is motivated by the study of transmission patterns of tuberculosis. We will shortly describe why the cluster size distribution is interesting in this context. Throughout the paper the terminology is motivated by this application. However, the related statistical problems are quite general and may occur in many different applications.

One of the major problems in tuberculosis epidemiology is that an infected person may not develop disease until decades after infection and the become infectious. In the mean the infection is said to be "dormant". Thus a new case of tuberculosis may be someone who was infected recently (on the order of months to a year), or someone who was infected long ago. In order to study the dynamics of transmission of tuberculosis, old and the recent infections need to somehow be identified. Using DNA fingerprinting techniques, usually based on RFLP patterns of IS6110-associated DNA polymorphism, it is possible to classify isolates from tuberculosis patients. Isolates with identical fingerprints form clusters of individuals. Assuming that the "fingerprints" are stable, persons with isolates that have differing fingerprints cannot be assumed to have infected each other. If, as assumed by Small et al. (1994), a cluster contains a unique primary source, the clustering of isolates carries information of the relation between reactivated infections and recent infections. If this is the case the proportion of all tuberculosis cases that are secondary cases can be calculated. This measure and a few other statistics based on the observed

cluster size distribution has been considered in a number of recent studies of tuberculosis. Methodological problems connected with the analysis of clustering data has been discussed by e.g., Glynn et al. (1999), Godfrey-Faussett (1999), Murray and Alland (2002), and Svensson (2002)).

In most studies it is not possible to classify all infectious persons with regard to fingerprint pattern. This may be due to difficulties to obtain isolates from all infectious persons or to failure to obtain a useful RFLP pattern. We will suggest a method to estimate the real cluster size distribution from the observed cluster sizes. The method is based on the assumption that the isolates that are not observed or classified can be regarded as a random sample of all isolates, and that the sampling proportion is known. This problem has also been addressed by Glynn et al. (1999). A different method has been suggested by Murray and Alland (2002).

2. Notation

Let

$$\begin{aligned} G_n &= \sharp\{\text{clusters of size n}\}, \\ H_i &= \sharp\{\text{clusters with i observed isolates}\}, \\ H_{in} &= \sharp\{\text{number of clusters of size n where i isolates are observed}\}. \end{aligned} \tag{2.1}$$

Obviously

$$H_i = \sum_{n \geq i} H_{in},$$

and

$$G_n = \sum_{i \leq n} H_{in}.$$

The probability to observe i isolates in a cluster of size n is

$$p_{in} = \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1-p)^{n-i}.$$

According to the assumptions

$$H_{in} \sim \text{Bin}(n, p),$$

and

$$\text{E}(H_i) = \sum_{n \geq i} p_{in} G_n.$$

Usually the cluster size distribution is summarized in a few cluster measures. In tuberculosis epidemiology two such measures are the proportion of recent infections and the proportion of clustered isolates. An isolate is said to be clustered if it belongs to a cluster with more than one member, i.e., if it is not unique. In the literature (cf. Small et al. (1994) and Glynn et al. (1999)) they are referred to as the $(n-1)-$ and the $n-$ measure respectively.

The are defined as

$$C_{n-1} = \frac{\sum_i (i-1)G_i}{\sum_i iG_i}, \tag{2.2}$$

and

$$C_n = \frac{\sum_i iG_i - G_1}{\sum_i iG_i}. \tag{2.3}$$

Observe that C_{n-1} has a one-to-one relation with the mean cluster size

$$M = \frac{\sum_i iG_i}{\sum_i G_i}, \tag{2.4}$$

since

$$C_{n-1} = 1 - \frac{1}{M}.$$

C_n has a one relation to the proportion of unique isolates

$$U = \frac{G_1}{\sum_i iG_i}, \tag{2.5}$$

since

$$C_n = 1 - U.$$

If these measures are calculated using the observed cluster sizes rather than the true cluster sizes the resulting estimates will be biased. One way to remove the bias is to estimate the actual number of clusters of different sizes, i.e., $G = (G_1, G_2, \ldots)$ from the observed sizes $H = (H_1, H_2, \ldots)$ and to use these estimates when calculating the clustering measures.

3. A moment estimate

A straightforward estimate of the true cluster size distribution is based on the moment relation given above. The true cluster sizes are estimated as solutions to the linear relation

$$H = PG, \tag{3.1}$$

where P is a matrix with elements p_{in}.

Since the matrix P is triangular it always has an inverse. Of course, the solution, H of 3.1 need not be a vector of non–negative integers. This can be corrected, by rounding of the solutions to the nearest integer. However, the moment estimator has other drawbacks. The solution will give no cluster with size larger than the largest observed size. The moment estimator does not take the size of the sample (the number of isolates or the number of clusters) into account. The estimator is inadmissible, in the sense that the solution may contain negative numbers.

4. An EM–algorithm

An ML-estimate seems to be an alternative to the moment estimate discussed above. It is rather simple to get an expression of the likelihood of the observations treating the unknown G as a parameter. However, the ML equations have no explicit solutions and finding the maximum of the likelihood will cause numerical problems. For this reason we will suggest an algorithm which will provide an estimate that should have approximately the same properties as the ML-estimate.

First assume that G_i, $i = 1, 2, \ldots$ are independent random variables that are Poisson distributed with intensity λ_i. This implies that H_{in} are also Poisson distributed with intensities $p_{in}\lambda_i$ and, which is important, are all independent. The likelihood for the observations H_{in}, $i = 1, \ldots, n$, $n = 1, 2, \ldots$ has a very simple structure. The problem is to derive an ML-estimate based on observations of $H_1, H_2, \ldots$ only. Such a solution may be found with the use of an EM-algorithm. The algorithm is defined in the following way:

1. Guess the values of that are consistent with the observations, i.e., $\sum_{n \geq i} H_{in} = H_i$, $i = 1, 2, \ldots$.
2. Find the ML-estimates of λ_i, $i = 1, 2, \ldots$, based on the assumed values of H_{in}. These estimates are $\tilde{\lambda}_n = \sum_{i=1}^{n} H_{in}/(1 - (1-p)^n)$. (This is the M-step).
3. Generate new guesses $H_{in} = H_i p_{in} \tilde{\lambda}_n / \sum_{m \geq i} p_{im} \tilde{\lambda}_m$. (This is the E-step).
4. Repeat the procedure from step 2 until convergence.

ML-estimates of the interesting measures can now be derived from the estimates $\hat{\lambda}_i$, $i = 1, 2, \ldots$. They are derived as

$$\hat{C}_{n-1}(p) = \frac{\sum_i (i-1)\hat{\lambda}_i}{\sum_i i\hat{\lambda}_i},$$

and

$$\hat{C}_n(p) = \frac{i\hat{\lambda}_1 - \hat{\lambda}_1}{\sum_i i\hat{\lambda}_i}.$$

The argument p indicates that the estimates depend on the assumed ascertainment probability. The special case $p = 1$ corresponds to using the observed cluster sizes.

To apply the algorithm the dimension of the parameter vector λ has to be decided. In principle it would be an infinitely-dimensional parameter. However we would be rather safe to assume that it is less than, e.g., $d_H/p + 4\sqrt{d_H(1-p)/p^2}$, where d_H is the largest observed cluster size.

The sample distribution of the estimates can be approximated using a (parametric) bootstrap technique. The following procedure is applied

1. From the observed cluster sizes $H = (H_1, H_2, \ldots)$ the ML-estimates $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2, \ldots)$ are derived under the assumption of a known value of p.

2. The true cluster sizes are estimated from $\hat{\lambda}$, by choosing an integer valued vector $\hat{G}$ so that $\max_i \mid \sum_{j \leq i} (\hat{G}_j - \hat{\lambda}_j) \mid$ is minimized.

3. A new set of "observations" $\hat{H}_i = \sum_n \hat{H}_{in}$, $i = 1, 2, \ldots$, are simulated by generating independent Multinomial distributed random vectors $(H_{0n}, \ldots, H_{nn})$ with parameters $(\hat{G}_n, p_{0n}, \ldots, p_{nn})$.

4. ML-estimates of the mean cluster size and the proportion of unique isolates are derived from the simulated (observed) cluster sizes.

5. Steps 3 and 4 are repeated until a sufficient number of replicates of the estimates is obtained.

An alternative way of deriving properties of the estimators would be to apply asymptotic theory. However, it should be observed that the situation is non-standard since the underlying parameter space, i.e., the λ's, has infinite dimension. Still it should be possible to verify that the distribution of the estimators are asymptotically normal. An formula for asymptotic variance of the estimator can also be calculated.

5. Numerical examples

We will illustrate the use of the estimation method with three numerical examples.

In the first example we will start with a hypothetical known cluster size distribution $G = (G_1, G_2, \ldots)$. From this distribution observed cluster size distributions are simulated. Simulations are made with different ascertainment probabilities. The method is then applied to the simulated distributions. The purpose is to investigate if the estimates come close to the values of the cluster measures, which in this case are known.

The second example is based on an observed cluster size distribution. In this example the ascertainment probability is not known. We assume different values of the ascertainment probability and use the method to derive estimates of the unknown cluster measures. The precision of the estimates are evaluated using bootstrap simulation as described above. From the calculations is it seen an estimate based on the observed cluster sizes will differ from estimates obtained after correcting for incomplete ascertainment. It is seen that the random variation caused by the fact that only a random sample of the isolates are fingerprinted is , at least, of the same magnitude as the bias.

In the third example is also based on observed cluster sizes. here only a (known) fraction of the tuberculosis patient are fingerprinted.

Table 5.1
True cluster sizes used in the hypothetical example

Cluster size	# of clusters	of isolates
1	64	64
2	32	64
3	16	48
4	8	32
5	4	20
6	2	12
7	1	7
all	127	247

Table 5.2
Mean, quantiles and standard deviation for estimates $\hat{C}_{n-1}(p)$ and $\hat{C}_n(p)$ in 1000 simulations with different ascertainment probabilities

p	$\hat{C}_{n-1}(p)$ mean	5 % percentile	95 % percentile	st.d	$\hat{C}_n(p)$ mean	5 % percentile	95 % percentile	st.d.
1.0	0.486	-	-	-	0.741	-	-	-
0.9	0.485	0.466	0.501	0.011	0.739	0.708	0.766	0.017
0.7	0.484	0.446	0.520	0.023	0.739	0.668	0.809	0.043
0.5	0.484	0.413	0.554	0.042	0.739	0.603	0.896	0.091

5.1 *Hypothetical distribution*

The calculations are based on the true cluster given in table 5.1. The mean cluster size equals 1.945 (= 247/127), and the proportion of unique isolates isolate is 0.259 (= 64/247). This implies that the true values of $\hat{C}_{n-1} = 0.486$ and $\hat{C}_n = 0.741$.

For each of the ascertainment probabilities $p = 0.5, 0.7$, and 0.9 1000 simulations of observed cluster sizes have been made. The estimates of the cluster measures are from the EM-algorithm suggested above are given in the table 5.2. The estimates seems to have a small bias. As can be expected the variation of the estimate increases as the ascertainment probability decreases. It can be shown that the expected values of C_{n-1} and C_n calculated from observed cluster sizes are increasing in p (cf. Glynn et al. (1999) and Svensson (2002)). If the ascertainment probability is known this bias can be removed, at least in this example.

Table 5.3
Observed cluster sizes for 473 Tuberculosis patients in San Francisco

Cluster size	# of clusters	of isolates
1	282	282
2	20	40
3	13	39
4	4	16
5	2	10
8	1	8
10	1	10
15	1	15
23	1	23
30	1	30
all	326	473

5.2 *San Francisco data*

The cluster sizes observed for the 473 San Francisco patients with tuberculosis analyzed by Small et al. (1994) are given in table 5.3. This cluster sizes gives the observed mean cluster size 1.45 (= 473/326), and the proportion of clusters with only one isolate is 0.87 (= 282/326). This implies that $\hat{C}_{n-1} = 0.31$ and $\hat{C}_n = 0.40$.

Table 5.4 gives the estimates of these parameters for different values of p. It turns out that the estimates of two proportion varies with the ascertainment probability assumed. As should be suspected, the estimates decreases as function of p. However the effect is rather moderate. Both C_{n-1} and C_n are proportions and it should be fair to evaluate the change using an oddsratio. For C_{n-1} the oddsratio is ≈ 0.77 and for C_n it is ≈ 0.87 when the ascertainment drops from $p = 1$ to $p = 0.5$.

The statistical properties of the estimates are investigated with the bootstrap method described above. The results are represented in table 5.5. As could be suspected the variability in the estimates increases as p decreases. The standard deviation of the estimates is approximately of the same size as the bias.

5.3 *South African data*

In a survey of South African gold miners 438 tuberculosis cases were identified (Godfrey-Faussett et al. (2000)). Fingerprints that could be used for dividing the isolates into clusters were only obtained from 371 og these isolates. The data are presented in table 5.6. In this case the part of the ascertainment that is due to problems of fingerprinting existing isolates is known. Of the 438 isolates 371 ,i.e. $371/438 \approx 0.847$ are used to define clusters. We will assume that the isolates are fingerprinted independently of each other and with the same probability, i.e., that the isolates that are divided into clusters is a random sample

Table 5.4

estimates of the cluster measures for the San Francisco data

p	$\hat{C}_{n-1}$	$\hat{C}_n$
1.0	0.311	0.404
0.9	0.318	0.410
0.8	0.329	0.418
0.7	0.341	0.425
0.6	0.354	0.431
0.5	0.369	0.438
0.4	0.384	0.444

Table 5.5

Mean, quantiles and standard deviation for 100 bootstrapped ML-estimates for $\hat{C}_{n-1}(p)$ and $\hat{C}_n(p)$, San Francisco data

p	$\hat{C}_{n-1}(p)$ mean	5 % percentile	95 % percentile	st.d	$\hat{C}_n(p)$ mean	5 % percentile	95 % percentile	st.d.
0.9	0.315	0.303	0.326	0.007	0.406	0.387	0.421	0.010
0.7	0.336	0.313	0.357	0.013	0.423	0.389	0.455	0.018
0.5	0.381	0.353	0.415	0.018	0.453	0.412	0.494	0.024

Table 5.6

Result of fingerprinting of 438 Tuberculosis patients in a South African mining community

Cluster size	# of clusters	of isolates
1	123	123
2	29	58
3	12	36
4	10	40
5	5	25
6	2	12
7	2	14
20	1	20
43	1	43
not observed		67
all	185	438

Table 5.7

Cluster measures based on observed cluster sizes and estimates with a 95 %, confidence interval based on ascertainment probability $p = 0.847$

cluster measure	based on observed cluster sizes	estimate	95 % confidence interval
C_{n-1}	0.501	0.523	[0.505, 0.543]
C_n	0.668	0.689	[0.665, 0.719]

of all isolates. Of course, this may not be the case since the possibility to obtain a good fingerprint may depend on properties of the isolate correlated to the fingerprint pattern.

In table 5.7 calculation of the cluster measures based on the observed cluster sizes and on estimates corresponding to $p = 0.847$ are given.

Acknowledgements

This work has been funded by The Bank of Sweden Tercentenary Foundation.

References

Glynn, J. R., Bauer, J., de Boer, A. S. et al. (1999). Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis. *International Journal of Tuberculosis and Lung Disease* **3**, 1055–1060.

Godfrey-Faussett, P. (1999). Interpretation of cluster studies of tuberculosis. *Lancet* **353**, 427–8.

Godfrey-Faussett, P., Sonnenberg, P., Shearer, S. et al. (2000). Tuberculosis control and molecular epidemiology in a South African gold-mining community. *The Lancet* **356**, 1066–1071.

Murray, M. and Alland, D. (2002). Methodological problems in the molecular epidemiology of tuberculosis. *American Journal of Epidemiology* **155**, 565–571.

Small, P. M., Hopewell, P. C., Singh, S. P. et al. (1994). The Epidemiology of Tuberculosis in San-Francisco - a Population-Based Study Using Conventional and Molecular Methods. *New England Journal of Medicine* **330**, 1703–1709.

Svensson, A. (2002). Diversity indices for infectious strains. Research Report 2002:5, Mathematical Statistics, Stockholm University.