Mathematical Statistics
Stockholm University

Clusters of Mycobacterium tuberculosis isolates – remarks concerning the comparability between populations and studies

Åke Svensson

Research Report 2002:8

ISSN 1650-0377

Postal address:

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

Internet:

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

Clusters of *Mycobacterium tuberculosis* isolates – remarks concerning the comparability between populations and studies

Åke Svensson

University of Stockholm and Swedish Institute for Infectious Disease Control

October 24, 2002

Abstract

In many studies of tuberculosis DNA fingerprinting techniques is used to classify isolates from patients. The isolates is then been divided into clusters with identical fingerprints and various measures of the amount of clustering are calculated. One purpose is to estimate the proportion of cases that is due to recent transmission. It is shown that cluster measures, generally, depends on the rate of reactivation of dormant infections, the diversity of strains, the rate of transmission of infections and the duration of the study. The understanding of how these factors influences the measures is essential for understanding how to interpret and compare results from different studies and also how to compare clustering in subpopulations.

1. Introduction

Using DNA fingerprinting techniques it is possible to classify isolates from tuberculosis patients. The classification is usually based on RFLP patterns of IS6110-associated DNA polymorphism. The classification makes it possible to form clusters of isolated with identical fingerprints. Isolates with the same fingerprint can be expected to be related to the same transmission chain or, at least, assuming that the "fingerprints" are stable, the opposite should hold, i.e., persons with isolates that have differing fingerprints can not be assumed to have infected each other. If there is a large variation of distinguishable fingerprints in the population it is tempting to assume that all primary sources of infections have different fingerprint and that accordingly patients with isolates with identical fingerprints are infected by the same primary case.

This idea has been used to analyze which proportion of tuberculosis is due to recent transmission and which is caused be reactivated infections. Small et al. (1994) made a study of the clustering of isolates from patient with tuberculosis reported to the San Francisco Department of Public Health, Division of Tuberculosis Control, between January 1, 1991, and December 31, 1992. The idea was to use the fingerprinting to analyze importance of recent transmission. The proportion of "clustered" patients (defined as the patient whose isolates had fingerprints that were not unique) and the proportion of the active tuberculosis cases that were the result of recent infection were calculated. Here the number of recent infections is calculated as the number of patients minus the number of

clusters. This is based on the assumption that each cluster contains exactly one primary case, which is a reactivated dormant infection.

In Borgdorff et al. (2001) an attempt is made to distinguish the primary case in each cluster, assuming that there is a single primary case. Alland et al. (1994) present a similar study of the spread of tuberculosis in New York City. Recently a large number of investigations based similar ideas have been carried through. Methodological problems related to the use of DNA fingerprinting has been discussed by Godfrey-Faussett (1999) and Murray and Alland (2002). They address a number of questions related to the interpretation of different measures of clustering, of comparing sub-populations and analyzing risk factors for recent transmission.

The purpose of the present paper is to discuss how some, more or less implicit, assumptions influence the results and the interpretations of the data. We will focus on

- The consequences on "clustering" if the fingerprints of the primary cases are not distinct. This implies that in a cluster of patients with identical fingerprints there may be several primary cases.
- The effect of the duration of the study, the reactivating intensity and the transmission rate in the population under study.
- The possibilities to compare "clustering" in different subpopulations.

Of course there are more assumptions, than those considered in this paper, that needs to be evaluated. It is, e.g., important that the fingerprints are reasonable stable in time. We will disregard the possibility that fingerprints changes during the study period. de Boer et al. (1999) has calculated a half-life time of 3.2 years for isolates from infectious patients (cf also Yeh et al. (1998)). This is of importance if the duration of the study is long. Vynnycky et al. (2001) consider the effect of such instabilities and also the effects of age related infectivity. Another important assumption is that an individual only can carry one strain at a time. There are some evidence that this may not be the case cf Yeh et al. (1999).

In section 2 we discuss the difference between transmission chains and clusters. Here the basic notations used in the paper are defined. A simple model for the reactivation of dormant infections and for recent spread is defined in section 3. The model is used, to analyze the impact of diversity and reactivating and transmission dynamics on different cluster measures. In section 4 measures obtained from different kinds of subpopulations is discussed. The relation between cluster measures in a subpopulation and the measures in the entire population is considered in section 5.

2. Clusters and transmission chains

We will distinguish between transmission chains and cluster of isolates. A transmission chain contains all cases derived from the same source, here called the primary case, via a chain of infections. A cluster of isolates is made up of all isolates with the same fingerprint

pattern. All isolates in the same transmission chain belongs to the same cluster. However, a cluster may consist of more than one chain. In general, we can not assume that there is a one-to-one correspondence between transmission chains and clusters.

The properties that are of real importance to describe the spread of infectivity are related to transmission chain rather than to clusters. Thus it is of interest to know the mean chain size (i.e. the mean number of isolates in the transmission chain) and the proportion of primary cases that are not attached to any secondary cases. If each cluster consists of only one transmission chain these measures correspond to the mean cluster size, and the proportion of unique isolates (i.e. the proportion of isolates that are not clustered).

2.1 *Measures and notation*

We will denote the number of clusters with i members by, G_i. The total number of isolates equals

$$I = \sum_i iG_i, \tag{2.1}$$

the number of clusters equals

$$K = \sum_i G_i, \tag{2.2}$$

and the number of unique isolates equals

$$U = G_1. \tag{2.3}$$

Several measures of the "amount of clustering" have been suggested. We will consider two such measures which are commonly used. In the paper by Glynn et al. (1999) they are referred to as derived by the n-method and the $(n-1)$-method. These measures were already defined by Small et al. (1994). A discussion in Murray and Alland (2002) illuminates the interpretation of these measures.

The first measure, C_n, describes the proportion of "clustered" isolates. An isolate is clustered if it belongs to a cluster with two or more members. Thus the proportion of "clustered" isolates is:

$$C_n = \frac{\sum_i iG_i - G_1}{\sum_i iG_i}. \tag{2.4}$$

This measure is closely related to the proportion of unique isolates

$$\frac{U}{I} = \frac{G_1}{\sum_i iG_i} = 1 - C_n. \tag{2.5}$$

A second measure, C_{n-1}, is natural to consider if we assume each cluster contains a unique primary source, i.e, all but one member in the cluster are secondary cases. Then the proportion secondary cases is:

$$C_{n-1} = \frac{\sum_i (i-1)G_i}{\sum_i iG_i}. \tag{2.6}$$

This measure is related to the mean cluster size

$$M = \frac{I}{K} = \frac{\sum_i iG_i}{\sum_i G_i} = \frac{1}{1 - C_{n-1}}, \tag{2.7}$$

since

$$C_{n-1} = 1 - 1/M.$$

3. Diversity and dynamics

3.1 *A simple model*

We will consider cases that occur during in a study of tuberculosis cases in a certain population. Assume that the study takes place during the time interval $(0, T)$. During this time we identify tuberculosis cases and fingerprint isolates. There is no reason to assume that the identified cases comes from primary sources that are reactivated during the time of the study. Our basic assumption is that each tuberculosis infection that is reactivated before the end of the study generates a (random) number of cases identified during the study. These cases may or may not include the primary case.

The clustering will depend on several basic parameters related to the population and the study. We will consider:

- The diversity of the strain distribution of (possible) primary cases,
- The reactivation rate at which dormant cases becomes infectious,
- The duration of the study.
- The speed which a transmission chain develops.

We will only consider incident cases, that is cases that are noted for the first time during the study period. Such cases may belong to transmission chains that have started before the study started. Clusters formed by incident cases do not, necessarily, include the primary (reactivated) cases. This should be kept in mind, since it invalidates the usual motivation for the clustering measure, C_{n-1}.

Figure 3.1 give a schematic illustration of how the cases included in the study are related to transmission chains and clusters.

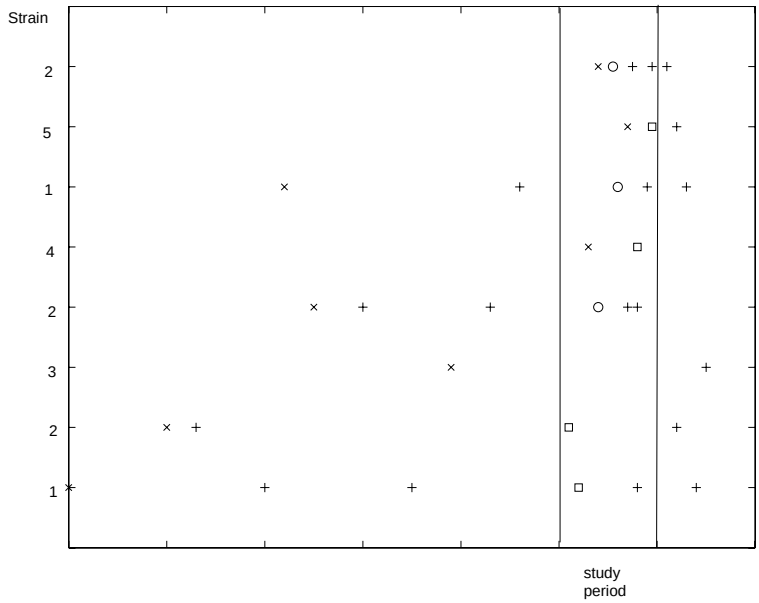

Figure 3.1. Illustration of observation plan. Each horizontal line refers to a transmission chain. x denote reactivated infections, $\square$ incident clusters, o incident chains, $+$ incident cases

3.1.1 Diversity The concept of diversity is used in ecology, genetics and economics to describe the relative abundance of subgroups in a population. It may concern the distribution of different species in the flora or fauna, the distributions of alleles at the same chromosomal locus of the distributions of individuals in income classes. Here our interest is focused on the distribution of fingerprints in the population of possible primary cases. In particular we will see how the cluster measures are related to the diversity in a population.

Suppose that are in total N (which may be a large number), of possible fingerprints. We consider two populations. In the first the distribution over the fingerprints are described by the proportions p_i, which is the proportion that has strain i, $i = 1, \ldots, N$. In the second population the proportions are $\tilde{p}_i$, $i = 1, \ldots, N$, Since we are only interested in the distribution of the strains and not on which of the strains is more or less common we may assume that the proportions are ordered within each population, i.e., $p_1 \geq p_2 \geq \ldots \geq p_N$ and $\tilde{p}_1 \geq \tilde{p}_2 \geq \ldots \geq \tilde{p}_N$. With this ordering the index i may not relate to the same strain in the two populations. It is natural to say that the strain distribution is more concentrated in the first population (or equivalently more diverse in the second) if

$$\begin{aligned} p_1 &\geq \tilde{p}_1, \\ p_1 + p_2 &\geq \tilde{p}_1 + \tilde{p}_2, \\ p_1 + p_2 + p_3 &\geq \tilde{p}_1 + \tilde{p}_2 + \tilde{p}_3, \\ &\ldots, \\ &\ldots \end{aligned} \tag{3.1}$$

A function ϕ is said to be Schur-convex if

$$\phi(p_1, \ldots, p_N) \geq \phi(\tilde{p}_1, \ldots, \tilde{p}_N)$$

if the distribution $(\tilde{p}_1, \ldots, \tilde{p}_N)$ is more diverse than the distribution $(p_1, \ldots, p_N)$, (cf Hardy et al. (1934) and Marshall and Olkin (1979)). If the function ϕ has the form

$$\phi(p_1, \ldots, p_N) = \sum_i f(p_i),$$

then it is Schur-convex if f is convex for $p \leq 1/2$.

3.1.2 Rate of reactivation We will assume that new transmission chains are started according to a (time homogeneous) Poisson process with intensity π. The probability that an observed chain carries the i'th strain is equal to the proportion, p_i, of dormant strains of type i. The assumption implies that all dormant cases are equally likely to be reactivated. Assume that we compare two populations, one with reactivating intensity π and the other with reactivating intensity $\tilde{\pi}$. We will say that the first of these populations has a higher reactivation if $\pi > \tilde{\pi}$.

3.1.3 Duration of the study The number of tuberculosis cases identified during the study that is caused by a certain reactivated case will depend on when the chain is initiated and how the chain develops in time. Let $X_s(0, T)$ denote a random variable that counts the number of cases identified during the study that are part of a transmission chain started by a infection that is reactivated at time s ($s \leq T$). According to this assumption the transmission dynamics does not depend on the strain involved. Of course, the number of cases identified will increase with the duration of the study.

3.1.4 Rate of transmission The rate at which incident cases occurs depends on the distribution of the random variables $X_s(0, T)$. When comparing two populations we will say that the transmission rate is faster in population A than in population B if for any s and T the distribution of $X_s(0, T)$ is stochastically larger in population than in population B. A random variable Z is said to be stochastically larger than $\tilde{Z}$ if $\mathrm{P}(Z \geq k) \geq \mathrm{P}(\tilde{Z} \geq k)$ for all k.

3.2 *Measures of clustering*

3.2.1 Basic statistics The first statistics that appears in an analysis are the number of isolates, the number of clusters and the number of unique clusters.

According to the model formulated above. The expected number of tuberculosis cases included in the study is:

$$\mathrm{E}(I) = \pi \int_{-\infty}^{T} E(X_s(0,T))ds. \tag{3.2}$$

This number will of course increase with T and π. It does not depend on the strain diversity. It will be larger the higher the rate of transmission is. As a function of T the expected number of (incident) isolates is a linearly increasing function. This implies that it is proportional to T, i.e.,

$$\mathrm{E}(I) = \rho\pi T.$$

The expected number of observed clusters is

$$\mathrm{E}(K) = \sum_j \left(1 - \exp(-\pi p_j \int_{-\infty}^{T} P(X_s(0,T) > 0)ds)\right). \tag{3.3}$$

Observe that this number depends the reactivation rate, the diversity of dormant strains and on the duration of study. The expected number of clusters will increase with π and T. Since the function defined by (3.3) is Schur-concave (cf Svensson (2002)) the expected number of clusters are larger the more diverse the strain distribution is.

The expected number of unique isolates is

$$\mathrm{E}(G_1) = \sum_j \pi p_j \int_{-\infty}^{T} P(X_s(0,T) = 1)ds \exp(-\pi p_j \int_{-\infty}^{T} P(X_s(0,T) > 0)ds). \tag{3.4}$$

Also this number depends on all factors considered in the model. However, there are no simple inequality relations.

3.2.2 C_{n-1} and C_n Number the incident cases occurring after time $t = 0$ consecutively by $i = 1, 2, \ldots$. Some of these cases will incident chains, i.e. be the first observed belonging to a particular transmission chains. Let D_i denote the number of incident chains and K_i denote the number of incident strains (clusters) among the i first cases. The strains attached to each chain is a mark which is given with probability equal to the proportion of dormant strains of that type. This marking does not influence the flow of incident cases or the flow of incident chains.

Now

$$\mathrm{E}(K_i \mid D_i) = \sum_j (1 - (1 - p_j)^{D_i}). \tag{3.5}$$

The function of the right-hand side of the equality is a Schur-concave function. This the expected number of observed clusters among the first i observed cases are larger the more diverse the strain distribution is. Thus the expectation of K is an increasing function of the diversity.

We can also show $\mathrm{E}(K_{i+2} - K_{i+1}) \leq \mathrm{E}(K_{i+1} - K_i)$ for all i. This implies that the expectation of K_i/i is a decreasing function in i. The more incident cases that are observed the larger the expected value of C_{n-1} will be. Thus $\mathrm{E}(C_{n-1})$ is increasing both with the reactivation rate and the duration of the study.

The inequalities we have obtained are all related to C_{n-1}. There seems not to be any obvious inequalities for C_n.

4. Measures of clustering in a subpopulation

It is sometimes of interest to restrict the study to a subpopulation of a larger population. The considerations made above are still valid if $X_s(0,T)$ ar interpreted as the number of cases in the subpopulation that turns up in the study. Of course, the distribution of these numbers depend on the interaction between the subpopulation and the rest of the entire population.

In this section we will analyze how the cluster measures for a subpopulation are related to the corresponding measures from the entire population. To do this we have to consider how the subpopulation is formed. We will here only consider three cases.

The first case concerns a totally random subpopulation, i.e., we assume that the subpopulation is a random sample of the complete population. This means that the subpopulation is formed by selecting each member of the population independently of each other with the same probability. Such a subpopulation can, e.g., be the responders in survey.

Since it is well known that a large part of the infectious spread is local, e.g. within a family, it is natural to assume that a considerable part of the spread takes place within the subpopulation. The second case relates to a model where the possibility is considered.

The third example is a subpopulation which can not generate primary cases in itself. A typically example of such a subpopulation are all children.

4.1 *Random subpopulation*

We assume that the subpopulation is formed from the total population as a random sample. The selection is assumed to be made with replacement, i.e., there is a certain individual belongs to the subpopulation with a certain probability p. An alternative model is that the subpopulation is a random sample of m individuals out of the n individual in the total population, i.e., a selection without replacement, where $p = m/n$. An important feature of this kind of subpopulation is that the individuals interacts with other individuals in the same way regardless if they are members in the subpopulation or not.

Table 4.1

Number of secondary cases in a transmission chain depending on which subpopulation the primary case comes from

Primary case	secondary cases in S	secondary cases in non-S
S	X_1	X_2
non-S	Y_1	Y_2

After the selection we will form the clusters of infected with isolates of the same type in the subpopulation. Observe that it is not certain that the primary case belongs to the population. The cluster measures obtained in this way will be denoted by $C_n(p)$ and $C_{n-1}(p)$.

In order to see how the selection procedure influences the measures of clustering we will first consider the effect of subtracting exactly one isolate from the population. If the subtracted isolate is unique isolate then C_{n-1} changes from $1-I/K$ to $1-(I-1)/(K-1)$ and if it is a non-unique isolate it changes from $1-I/K$ to $1-(I-1)/K$. If the individual is chosen at random the first of these changes will take place with probability G_1/I and the second with probability $1-G_1/I$. A simple calculation yields that the expected value of the change is negative. Repeating this procedure in step by step we see that the expected value of C_{n-1} decreases as the more and more individuals are subtracted. Thus a random selection of m out of n individuals without replacement will yield a lower expected value the smaller m is. From this we can conclude that $\mathrm{E}(C_{n-1})$ is an increasing function of p.

Similarly subtracting one isolate at random will change C_n from $1-G_1/I$ to $1-(G_1-1)/(I-1)$ with probability G_1/I, to $1-(G_1+1)/(I-1)$ with probability $2G_2/I$, and to $1-G_1/(I-1)$ with probability $1-(G_1+2G_2)/I$. Also this change has a negative expectation. Thus $\mathrm{E}(C_n)$ is an increasing function of p.

The last fact has been observed by Glynn et al. (1999) for a subpopulation that has been randomly selected with a fixed size, i.e. the subpopulation has been selected without replacement rather than with fixed probability.

4.2 *Clustering within the subpopulation*

We will consider a division of the entire population into two subpopulations S and N. N are all individuals that do not belong to S. We will assume that a primary case may cause a random number of secondary cases in both subpopulations. The probability law for these bivariate random vector depends on from which population the primary case comes. Table 4.1 gives the notation for these random variables.

It could be expected that X_1 and X_2 (as well as Y_1 and Y_2) are positively correlated. Let V_S be the number of members in the subpopulation S in a chain and V_N be the number of members outside S. Furthermore let p be the probability that a primary case in the population comes from subpopulation S and let $q=1-p$. Elementary calculations yields:

$$
\begin{aligned}
\mathrm{E}(V_S) &= p + p\mathrm{E}(X_1) + q\mathrm{E}(Y_1), \\
\mathrm{Var}(V_S) &= pq(1 + \mathrm{E}(X_1) - \mathrm{E}(Y_1))^2 + p\mathrm{Var}(X_1) + q\mathrm{Var}(Y_1), \\
\mathrm{Cov}(V_S, V_N) &= -pq(1 + \mathrm{E}(X_1) - \mathrm{E}(Y_1))(1 + \mathrm{E}(Y_2) - \mathrm{E}(X_2)) + \\
&\quad p\mathrm{Cov}(X_1, X_2) + q\mathrm{Cov}(Y_1, Y_2).
\end{aligned} \tag{4.1}
$$

It is also possible to calculate the conditional distribution of the number of members from subpopulation S in a cluster of size n.

$$
\begin{aligned}
\mathrm{E}(V_S \mid V_S + V_N = n) &= np + (q\mathrm{E}(Y_1 \mid Y_1 + Y_2 = n) - p\mathrm{E}(X_2 \mid X_1 + X_2 = n)) \\
\mathrm{Var}(V_S \mid V_S + V_N = n) &= p\mathrm{Var}(X_1 \mid X_1 + X_2 = n) + q\mathrm{Var}(Y_1 \mid Y_1 + Y_2 = n) + \\
&\quad pq(\mathrm{E}(X_1 \mid X_1 + X_2 = n) - \mathrm{E}(Y_1 \mid Y_1 + Y_2 = n))^2
\end{aligned} \tag{4.2}
$$

Assuming that the potential to reactivate and spread the infection is the same in the subpopulation as in the entire population it is natural to assume that

- p equals the proportion of the population that belongs to S,
- $X_1 + X_2 = Y_1 + Y_2$ in distribution, i.e., the distribution of the number of secondary cases is the same independent of which population the primary case comes from, and
- $p\mathrm{E}(X_2) = q\mathrm{E}(Y_1)$, i.e., the expected spread outside of the subpopulation from which the primary case comes from is proportional to the size of population.

We will call a subpopulation with this property balanced (within the population). Observe that a random subpopulation is balanced.

Another example, of a balanced subpopulation, occurs if we assume that each primary case gives rise to a random number, Z_1, of cases exclusively in its own subpopulation and an independent random number, Z_2 which is distributed randomly in the entire population. Formally we can write

$$
\begin{aligned}
X_1 &= Z_1 + \sum_{i=1}^{Z_2} \delta_i, \\
X_2 &= \sum_{i=1}^{Z_2} (1 - \delta_i), \\
Y_1 &= \sum_{i=1}^{Z_2} \delta_i, \\
Y_2 &= Z_1 + \sum_{i=1}^{Z_2} (1 - \delta_i),
\end{aligned} \tag{4.3}
$$

where δ_i, $i = 1, \ldots, Z_2$ are independent random indicators such that $\delta_i = 1$ with probability p, and Z_1 and Z_2 are independent random variables.

In a balanced population

$$\mathrm{E}(I_S) = p\mathrm{E}(I). \tag{4.4}$$

If there is a one-to-one correspondence between chains and clusters then

$$\mathrm{E}(K_S) = (p + q\mathrm{P}(Y_1 > 0))\mathrm{E}(K). \tag{4.5}$$

If a cluster can contain isolates from more than one transmission chain the rate of occurrence of new chains and the diversity of strains has to be considered. According to the observation plan the number of incident chains will be Poisson distributed with the mean, λ, that depends on reactivation rate and the duration of the study. The number of incident chains with strains i will be Poisson distributed with mean λp_i. Then

$$\mathrm{E}(K_S) = \sum (1 - \exp(-\lambda p_i(p + q\mathrm{P}(Y_1 > 0))). \tag{4.6}$$

When considering the number of unique clusters it is important to distinguish between the number of clusters with exactly one member in the subpopulation S (denoted by U_S) and the number of clusters that is represented by one unique isolate in the entire population which belongs to S (denoted by $\tilde{U}_S$).

$$\mathrm{E}(\tilde{U}_S) = p\mathrm{E}(U). \tag{4.7}$$

If there is a one-to-one correspondence between chains and clusters then

$$\mathrm{E}(U_S) = p\mathrm{E}(U) + q\mathrm{P}(Y_1 = 1)\mathrm{E}(K). \tag{4.8}$$

The general expression is

$$\mathrm{E}(U_S) = \sum \lambda p_i(p + q\mathrm{P}(Y_1 = 1)) \exp(-\lambda p_i(p + q\mathrm{P}(Y_1 = 1)). \tag{4.9}$$

Approximating the mean cluster size with the ratio between the expectation of I_S and K_S, we see that the mean number of cases in a cluster with at least one member from the subpopulation S increases with p. The same holds for the mean number of members from S in clusters with at least one member from S. Since C_{n-1} is a function of the mean cluster size it will approximately have the same monotonicity properties.

Approximating C_n for the subpopulation in the same way

$$C_n \approx 1 - \frac{\mathrm{E}(U_S)}{\mathrm{E}(I_S)} \tag{4.10}$$

there is no monotonicity related to p. However, observe that

$$\frac{\mathrm{E}(\tilde{U}_S)}{\mathrm{E}(I_S)} \tag{4.11}$$

does not depend on p. Thus with a balanced subpopulation

$$\frac{\mathrm{E}(\tilde{U}_S)}{\mathrm{E}(I_S)} = \frac{\mathrm{E}(\tilde{U}_N)}{\mathrm{E}(I_N)} = \frac{\mathrm{E}(\tilde{U})}{\mathrm{E}(I)}. \tag{4.12}$$

Table 5.1
Number of unique and clustered isolates in two subpopulations

	in S	in N	total
unique isolates	$\tilde{U}_S$	$\tilde{U}_n$	U
clustered isolates	$I_S - \tilde{U}_S$	$I_N - \tilde{U}_n$	$I - U$
all isolates	I_S	I_N	I

Table 5.2
The expected number of unique and clustered isolates in two balansed subpopulations
$W = \mathrm{P}(Z > 0) + \mathrm{E}(Z)$

	in S	in N	total
unique isolates	$\lambda p \mathrm{P}(Z = 0)$	$\lambda q \mathrm{P}(Z = 0)$	$\lambda \mathrm{P}(Z = 0)$
clustered isolates	$\lambda p W$	$\lambda q W$	λW
all isolates	$p\lambda(1 + \mathrm{E}(Z))$	$q\lambda(1 + \mathrm{E}(Z))$	$\lambda(1 + \mathrm{E}(Z))$

4.3 *Innocent subpopulation*

An innocent population is a population that can not include any individuals which can serve as primary cases. This is a special case of the situation treated in previous subsection, when $p = 0$. A typical such population is the population of small infants.

The formulas derived above are valid also in this situation

5. Test if a subpopulation is balanced

A division of the population into two parts, S and N, and a division of the unique and clustered isolates results in a two by two table:

According to the assumptions made the number of (incident) chains is Poisson distributed. Let its mean be denoted by λ. Then according to the calculations made above the expected number of unique and clustered isolates are given by table 5.2. Here Z is the number of secondary cases, which if the population is balanced have the same distribution regardless of which population the primary cases belongs to. If the populations are balanced the cross-product ratio

$$Q = \frac{\tilde{U}_S(I_N - \tilde{U}_N)}{U_N(I_S - \tilde{U}_N)} \tag{5.1}$$

should be, according to the calculations made above, be close to 1. It is possible to base a test of the hypothesis that the populations are balanced on this approximation.

Asymptotic theory and a Taylor expansion of the logarithm in (5.1) yields that if there are many transmission chain Q is asymptotically normal distributed with mean 0 and variance v^2 which can be approximated by $\tilde{v}^2$, where

$$\tilde{v}^2 = \frac{1}{\tilde{U}_S} + \frac{1}{\tilde{U}_N} + \frac{1}{I_S - \tilde{U}_S} + \frac{1}{I_N - \tilde{U}_N} + d^2. \tag{5.2}$$

If $d^2 = 0$ then the variance is the same as in two-by-two contingency table for test of homogeneity. This implies that a usual test for if the odds ratio equals 1 in such tables (or equivalently a common χ^2-test) is relevant.

In general $d^2 \neq 0$. Tedious but trivial calculations yield that

$$d^2 = \frac{1}{\lambda(\mathrm{P}(Z > 0) + \mathrm{E}(Z))^2} \left(\frac{\mathrm{E}(V_S(V_S - 1))}{p^2} + \frac{\mathrm{E}(V_N(V_N - 1))}{q^2} - \frac{2\mathrm{E}(V_S V_N)}{pq} \right). \tag{5.3}$$

In case we have a randomized population $d^2 = 0$. However, if the structure is as described in (4.3) then $Z = Z_1 + Z_2$ and

$$d^2 = \frac{\mathrm{E}(Z_1(Z_1 + 1))}{pq\lambda(\mathrm{P}(Z > 0) + \mathrm{E}(Z))^2}. \tag{5.4}$$

In this case the use of a standard χ^2-test of the hypothesis that the populations are balanced will be conservative, unless Z_1 also is equals 0.

In case there is a one-to-one correspondence between chains and clustered d^2 may be estimated from the observations. The parameter λ is estimated by the number of observed clusters, $\mathrm{P}(Z = 0)$ by the proportion of clusters with only one isolate, p by the proportion of isolates (or clustered isolates) in subpopulation S. The other moments in the expression (5.3) are estimated by their corresponding empirical moments.

Acknowledgements

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

References

Alland, D., Kalkut, G. E., Moss, A. R. et al. (1994). Transmission of Tuberculosis in New-York-City - an Analysis by DNA-Fingerprinting and Conventional Epidemiologic Methods. *New England Journal of Medicine* **330**, 1710–1716.

Borgdorff, M. W., N.J.D, N., P.R.W, d. H. and D., v. S. (2001). Transmission of Mycobacterium tuberculosis Depending on the Age and Sex of Source Cases. *American Journal of Epidemiology* **154**, 934–943.

de Boer, A. S., Borgdorff, M. W., de Haas, P. E. W. et al. (1999). Analysis of rate of change of IS6110 RFLP patterns of Mycobacterium tuberculosis based on serial patient isolates. *Journal of Infectious Diseases* **180**, 1238–1244.

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.

Hardy, G., Littlewood, J. and Pólya, G. (1934). *Inequalities*. Cambridge University Press, Cambridge.

Marshall, A. and Olkin, I. (1979). *Inequalities: Theory of majorization and its Applications*. Academic Press, London.

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.

Vynnycky, E., Nagelkerke, N., Borgdorff, M. W. et al. (2001). The effect of age and study duration on the relationship between 'clustering' of DNA fingerprint patterns and the proportion of tuberculosis disease attributable to recent transmission. *Epidemiology and Infection* **126**, 43–62.

Yeh, R. W., de Leon, P., Agasino, C. B., Hahn, J. A., Daley, C. L., Hopewell, P. C. and Small, P. M. (1998). Stability of Mycobacterium tuberculosis DNA genotypes. *Journal of Infectious Diseases* **177**, 1107–1111.

Yeh, R. W., Hopewell, P. C. and Daley, C. L. (1999). Simultaneous infection with two strains of Mycobacterium tuberculosis identified by restriction fragment length polymorphism analysis. *International Journal of Tuberculosis and Lung Disease* **3**, 537–539.