ELSEVIER

Preventive Veterinary Medicine 36 (1998) 11–23

PREVENTIVE VETERINARY MEDICINE

Prevalence estimation under heterogeneity in the example of bovine trypanosomosis in Uganda

Dankmar Böhning[a], Matthias Greiner[b,*]

[a] *Institute for Social Medicine and Medical Psychology, Department for Epidemiology, Freie Universität Berlin, Fabeckstraβe 60-62, 14195 Berlin, Germany*

[b] *Institute for Parasitology and Tropical Veterinary Medicine, Department of Tropical Veterinary Medicine and Epidemiology, Freie Universität Berlin, Königsweg 67, 14163 Berlin, Germany*

Accepted 2 April 1998

Abstract

We examine variance estimators of a binomial parameter established under cluster sampling using data from a cross-sectional study of bovine trypanosomosis in Mukono County, Uganda. Fifty farms (referred to as clusters), were sampled with a total sample size of 487 cattle. Trypanosomes were found in 17.9% (87/487) of the total sample. The cluster-level (CL) prevalences were not homogeneously distributed. According to maximum-likelihood parameters established by mixture-distribution analysis, 18% of the cluster had 0% prevalence whereas 48% and 34% of the clusters could be allocated to subpopulations of clusters with mean prevalences 11.6% and 31.9%, respectively. We show that this form of heterogeneity invalidates the applicability of the Beta distribution as a model for the distribution of CL prevalences. Furthermore, we provide empirical evidence for a variance inflation due to heterogeneity (inflation factor 2.07) that exceeds the design-based variance inflation due to clustering alone (inflation factor 1.82). The variance inflation due to heterogeneity is given in a closed form so that the approach can be conveniently applied to survey data that involve cluster sampling under heterogeneity. © 1998 Elsevier Science B.V.

Keywords: *Trypanosomes* sp.; Cluster sampling; Variance inflation; Mixture distribution

1. Introduction

Sampling weights, stratification and clustering are important sampling-design features that must be considered for the analysis of observational studies. Sampling weights reflect the probabilities of selection – which may vary between different observations.

* Corresponding author. Tel.: +49 30 8108 2317; fax: +49 30 8108 2323; e-mail: mgreiner@vetmed.fu-berlin.de

PII S0167-5877(98)00076-2

Generally, the weight of a given observation is proportional to the inverse of the probability of being sampled. Stratification involves an independent sampling across predetermined strata which are mutually exclusive parts of the sampling frame. Clustering occurs if the sampled observations stem from clusters such as – in veterinary epidemiology – herds, litters or flocks. The effects of these sampling-design characteristics must not be ignored during analysis because they may affect parameter estimates (sampling weights) and variance estimates (stratification and clustering). Variance estimates tend to be smaller when established for individual strata than for the entire sample. In contrast, cluster sampling tends to inflate variances due to intracluster correlation. The relevance of the latter effect referred to as 'cluster effect' – for veterinary epidemiology has been discussed in detail by McDermott et al. (1994). Methods are available for the estimation of design-based (e.g. accounting for complex sampling) variances including the variance inflation derived from the intracluster correlation (Donner, 1993).

The parameter of interest in prevalence studies is the proportion of events. The primary sampling unit (PSU) in veterinary epidemiology is usually the herd, litter or flock, which we denote as 'cluster' from now on. Thus, the distribution of cluster-level (CL) prevalences becomes an issue. The methods proposed to account for cluster effects accommodate variability in the CL prevalences which is often modelled using the Beta distribution (e.g. Donald et al., 1994). The distribution shapes that can be modelled using the Beta distribution comprise the unimodal with zero-inflation (L-shape), peaked distribution and one-inflation (J-shape) and the unique bimodal distribution with a combined zero- and one-inflation (U-shape). However, we have empirical evidence for distribution types that reflect heterogeneity of CL prevalences and that cannot be handled by the Beta distribution model. In our example, we are concerned with the prevalence estimation under a cluster-sampling design (Section 2.1). We use an estimate of the intracluster correlation coefficient (ICC) to estimate a variance inflation factor that takes into account the cluster sampling design. We denote this procedure as 'ICC-approach' (Section 3.1). The underlying assumptions of the ICC-approach are explained using the Beta-binomial distribution model. A more flexible approach is derived that has its foundation in the empirical observation of distribution heterogeneity in the CL prevalences. We denote this approach as 'heterogeneity-approach' (Section 3.2). We provide a formal argument that the heterogeneity-approach is less biased than the ICC-approach in case of unequal cluster sizes (Section 3.3). We suggest using well-established methods of non-parametric mixture modelling as a natural way to proceed (Böhning et al., 1998) and to identify heterogeneity in the sample (Section 3.4). Finally, the practical relevance of this work is discussed (Section 4).

2. Material and methods

2.1. Example data set

The data were collected during a cross-sectional study in June 1994 in Mukono County, located in the southeastern part of Uganda. The sampling frame consisted of 187 dairy farms existing in the region (information from census April 1994) from which 50

Table 1
Prevalence estimation of bovine trypanosomosis in cattle sampled from 50 dairy farms in Mukono County, Uganda. Number of infected cattle (cases), sample size, prevalence and classification into one out of three subpopulations of farms identified by mixture analysis (data from June 1994, total sample size 487)

Farm	Cases	Sample size	Prevalence	Subpopulation	Farm	Cases	Sample size	Prevalence	Subpopulation
1	4	9	0.44	3	26	1	7	0.14	2
2	0	5	0	2	27	1	3	0.33	3
3	3	9	0.33	3	28	1	11	0.09	2
4	14	32	0.44	3	29	1	3	0.33	3
5	2	17	0.12	2	30	1	3	0.33	3
6	0	3	0	2	31	1	9	0.11	2
7	1	4	0.25	2	32	4	9	0.44	3
8	3	17	0.18	2	33	0	9	0	1
9	0	7	0	2	34	0	7	0	2
10	0	15	0	1	35	3	19	0.16	2
11	0	8	0	2	36	1	13	0.08	2
12	0	12	0	1	37	0	12	0	1
13	0	9	0	1	38	5	18	0.28	3
14	0	16	0	1	39	2	11	0.18	2
15	6	16	0.38	3	40	0	12	0	1
16	2	5	0.40	3	41	0	2	0	2
17	0	9	0	1	42	2	7	0.29	3
18	0	6	0	2	43	2	7	0.29	3
19	2	8	0.25	3	44	4	10	0.40	3
20	0	6	0	2	45	3	10	0.30	3
21	0	3	0	2	46	1	3	0.33	3
22	1	7	0.14	2	47	1	15	0.07	2
23	1	8	0.13	2	48	0	6	0	2
24	0	10	0	1	49	1	6	0.17	2
25	12	28	0.43	3	50	1	6	0.17	2

farms were selected at random, stratified on small (1–10 cattle), medium (11–30) and large (>30) farms. A total of 487 cattle was sampled. The prevalence of bovine trypanosomosis was established using parasitological techniques. We have not reckoned with a strong ICC in our data set based on earlier analyses (Greiner et al., 1997; wherein further details of this study are described). Biologically, ICC of bovine trypanosomosis – if present – can be thought of reflecting a similar exposure and disease management for animals stemming from one farm. The herd-level prevalences are listed in Table 1. The mean sample size of the 50 farms was 9.7. Fig. 1(A) shows the frequency-distribution histogram of CL prevalences. According to a formula suggested by Kairisto (1995), we arbitrarily selected the bin width (b=0.9 [min (s, IR/1.34)] $n^{-0.2}$) for construction of the histogram, where s, IR and n denote the standard deviation, the interquartile range and the number of clusters, respectively.

2.1.1. Computer software used

The one-factor analysis of variance (ANOVA; see Section 3.1) requires that the aggregated count data for CL prevalences are to be rearranged into binary data (infection

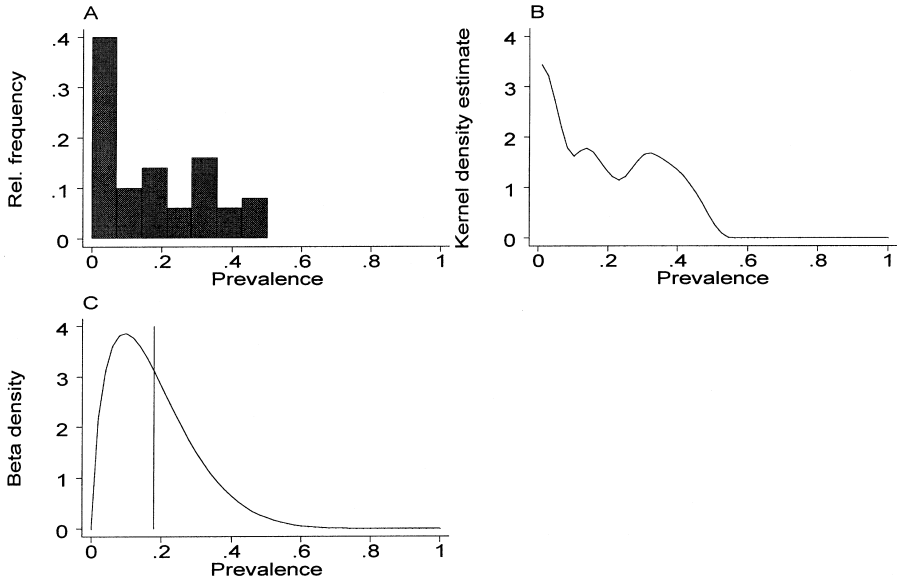

Fig. 1. Prevalence study of bovine trypanosomosis in Mukono County, Uganda (data from June 1994, total sample size N=487). (A) Observed frequency distribution of cluster-level (farm) prevalences. (B) Cosine kernel density trace of the same data. The shape suggests that the distribution is not unimodal but has three modes. (C) Beta distribution model of the same data using the total prevalence $\hat{p} = 0.1786$ (with mean cluster size 9.7) and the intracluster correlation coefficient $\hat{\rho} = 0.0939$ as coefficients. According to the Beta-distribution model, the mode and the expected value (vertical line at p=0.1786) do not coincide. Note that the x-scale of A and B refer to observed prevalences whereas the x-scale of C refers to expected prevalences under the Beta-distribution model.

status) at the element (animal level in our case) level (EL) with the cluster identification (farm number) as the grouping variable. A computer program (clusDATA) is available from the authors for this purpose (http://city.vetmed.fu-berlin.de/~mgreiner/clusDATA/clusdata.htm). The distribution of observed CL prevalences is displayed using a cosine kernel density estimation graph (short 'density trace'; Stata, Statacorp., 1997). The shape of a frequency-distribution histogram critically depends on the bin widths (b) of intervals and may not always adequately visualise the underlying distribution. This is especially true for small sample sizes, where the arbitrary selection of fixed-interval limits introduces a subjective element in the graphical representation of the data. The kernel density trace provides a solution to this problem. Similar to a moving average, this function evaluates the local probability density for each point basing on overlapping ranges of prevalence data. This technique is a simple non-analytical method of displaying a frequency distribution of a continuous variable (here, a proportion). The correlation between cluster (farm) size vs. prevalence and cluster sample size vs. prevalence were assessed by the coefficient of determination (squared Pearson's product-moment correlation) based on a linear-regression analysis with the prevalence as dependent variable (Stata; Statacorp., 1997).

3. Prevalence estimation in the presence of extrabinomial variation

3.1. Variance inflation due to the cluster effect

In prevalence studies of infectious diseases, the parameter of interest is usually the prevalence p. The prevalence is estimated as $\hat{p}=y/N$, where N denotes the total sample size and y the total number of infected animals out of N. Under the assumption of simple random sampling (srs), the naive expected variance using the binomial model is $\mathrm{VAR}_{\text{srs-bin}}(\hat{p}) = \hat{p}(1-\hat{p})/N$ (we omit the hat on VAR for all variance estimates). In our study, the prevalence estimate is $(\hat{p} = 87/487 = 0.1786)$ with $\mathrm{VAR}_{\text{srs-bin}}(\hat{p}) = 0.00030$. The latter estimate is not adequate with regard to the cluster sampling. The cluster effect results in a variance inflation because the number of independent observations is less than the denominator N. Thus, confidence limits for the prevalence are too small; the precision of the parameter estimate is overestimated (McDermott et al., 1994). As described by Brier (1980); Donner (1993); McDermott and Schukken (1994), we can analytically account for the cluster effect using the variance-inflation factor

$$c = 1 + \rho(\bar{n} - 1), \tag{1}$$

where ρ and $\bar{n}$ denote the intracluster correlation coefficient (ICC) and the average cluster sample size, respectively. The variance-inflation factor is equivalent to the so-called 'design effect' (*deff*) that expresses the ratio of the design-based variance and the variance expected under simple random sampling (Kish, 1965). Thus, the design-based inflated variance is given as $\mathrm{VAR}_{\text{ICC}}$=c $\mathrm{VAR}_{\text{srs-bin}*}$. This formula is widely used but we are not aware of a formal derivation of it. Therefore, a formal justification is presented in the appendix. If the sample was collected from a number of k clusters with the number of cases y_i and sample size n_i for the ith cluster we can estimate ρ from the data as

$$\hat{p} = (\mathrm{MSB} - \mathrm{MSW})/(\mathrm{MSB} + \mathrm{MSW}(\bar{n} - 1)) \tag{2}$$

where

$$\mathrm{MSB} = 1/(k-1)\sum_{i=1}^{k}(y_i - n_i\hat{p})^2/n_i$$

and

$$\mathrm{MSW} = 1/(N-k)\sum_{i=1}^{k}(n_i - y_i)/n_i$$

denote, respectively, the mean square between clusters and the mean square within clusters (Fleiss, 1981). For our study data, $\bar{n}$=9.7, MSB=0.2681, MSW=0.1334 and $\hat{p}$=0.0939 – resulting in the estimate for the variance-inflation factor $\hat{c} = 1.82$. We derive MSB and MSW from one-factor ANOVA with the cluster identification as the independent (grouping) variable and the event of infection on the individual-animal level (EL) as dependent variable. The 95% confidence interval (CI) for the prevalence $\hat{p}$ according to the Normal approximation $\mathrm{CI}(\hat{p}) = \hat{p} \pm 1.96(\mathrm{var}(\hat{p}))^{0.5}$ is $\mathrm{CI}_{\text{srs-bin}}(\hat{p}) =$ [0.1446, 0.2127] for the naive and $\mathrm{CI}_{\text{ICC}}(\hat{p}) = [0.1327, 0.2245]$ for the inflated

(design-based following the ICC-approach) variance estimation. According to the coefficient of determination (R^2), there was no evidence of a correlation between cluster (farm) sizes and prevalence (R^2=0.04) or between cluster sample sizes and prevalence (R^2=0.028).

3.2. Variance inflation due to heterogeneity

Variance inflation can have various causes. One of these can be seen in the fact that the parameter p of the Binomial distribution is varying in the population. This phenomenon is called (parameter) *heterogeneity*. Approaches differ in the way the heterogeneity distribution is modeled. A frequently employed model for the distribution of CL prevalences p is the Beta distribution (e.g. Donald et al., 1994):

$$\mathrm{B}(p/\alpha,\beta) = p^{(\alpha-1)}(1-p)^{(\beta-1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

where α and β are parameters and $\Gamma(\cdot)$ denotes the Gamma function. The Beta distribution has mean $\pi=\alpha/(\alpha+\beta)$ and variance $\tau^2=\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]=[1/(\alpha+\beta+1)]\,\pi(1-\pi)=\rho\pi(1-\pi)$. Note that π can be interpreted as the overall prevalence in the population. Also, the notation π has been used to distinguish it from the prevalence p under homogeneity. The link to the prevalence estimation under cluster sampling is through the parameters which can be established using as $\hat{\alpha}=\hat{p}/\hat{\rho}-\hat{p}$ and $\hat{\beta}=(1-\hat{p})/\hat{\rho}+\hat{p}-1$, where $\hat{p}$ and $\hat{\rho}$, respectively, are sample estimates of the prevalence π and the intracluster correlation coefficient ρ. In our case, $\hat{\alpha}$=1.724 and $\hat{\beta}$=7.928. Since $\hat{\alpha}>1$ and $\hat{\beta}>1$, we obtain a unimodal, peaked distribution density with the mode at $p=(\hat{\alpha}-1)/(\hat{\alpha}+\hat{\beta}-2)$=0.0946 (Fig. 1(C)). Since this distribution model is explicitly unimodal, it is not suitable for modelling multimodal distributions. In order to account for extrabinomial variation, the Beta distribution of p can be used instead of the fixed binomial parameter p in a discrete distribution model leading to the Beta-binomial distribution:

$$\text{Beta-binomial}(Y=y/\alpha,\beta,n) = \binom{n}{y}\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\alpha+\beta+n)}$$

A meaningful parameterisation is $\pi=\alpha/(\alpha+\beta)$ and the so called *index of aggregation* $\Theta=1/(\alpha+\beta)$. The latter can be linked to the intracluster-correlation coefficient through $\rho=\Theta/(1+\Theta)$ (Madden and Hughes, 1995). The variance of the Beta-binomial is $n\pi(1-\pi)[1+\rho(n-1)]$ which turns out to be the variance of the Binomial distribution $[n\pi(1-\pi)]$ multiplied by the variance-inflation factor given in Eq. (1). The Beta-binomial is an overdispersion model and equivalent to the classical Binomial if and only if $\rho=0$. In this case, the variance of the Beta distribution becomes zero and the variance of the Beta-binomial distribution becomes equal to the variance of the binomial.

Empirically, we are aware of distributions of CL prevalences that cannot be described by the Beta distribution. The latter model is inflexible in the sense that it can only accommodate peaked distributions (in case of $\alpha>1$ and $\beta>1$), L-shaped (i.e. zero-inflation incase of $\alpha>1\geq\beta$) and J-shaped (i.e. one-inflation in case of $\alpha\leq1<\beta$) and the unique bimodal case with U-shaped distribution (i.e. combined zero- and one-inflation in case of

α<1 and β<1). Distributions with more than two modes or bimodal distributions with modes others than zero and one cannot be accommodated. We also note that the unique case of bimodal distribution of the U-shape requires that $\hat{p} < \hat{\rho}/(1-\hat{\rho})$ and $\hat{\rho} > (1-\hat{p})/(2-\hat{p})$. The unimodal distribution model (Fig. 1(C)), however, is not in agreement with the empirical evidence for heterogeneity in our data. The frequency-distribution histogram (Fig. 1(A)) and – even more clearly – the density trace (Fig. 1(B)) suggests that three subpopulations of farms (clusters) exist, with mean prevalences of about 0%, 15% and 35%, respectively. Such a distribution cannot be modeled using the Beta distribution. In the following, we are interested to provide a more flexible, general approach for modelling heterogeneity.

For the sake of simplicity, we consider the Poisson approximation for the variance of the estimated prevalence $\mathrm{VAR}_{\text{srs-poi}}(\hat{p}) = \hat{p}/N$, which is equivalent to y/N^2. In our study, the naive (Poisson-based) variance is $\mathrm{VAR}_{\text{srs-poi}}(\hat{p}) = 0.00037$. In the case of cluster sampling with k clusters, the pooled estimator $\hat{p}_{\text{pool}}=(y_1+\ldots+y_k)/(n_1+\ldots+n_k)$ is the commonly used estimator of the overall prevalence p. The variance of $\hat{p}_{\text{pool}}$ under homogeneity is readily provided as $\mathrm{VAR}_{\text{srs-poi}}(phat_{\text{pool}})=(\hat{p}n_1+\ldots+\hat{p}n_k/(n_k+\ldots+n_k)^2=\hat{p}/N$, with $N=n_1+\ldots+n_k$. Note that this variance can be estimated by $(y_1+\ldots+y_k)/N^2$. In the case of heterogeneity, the variance of the total prevalence $\hat{p}_{\text{pool}}$ is inflated by a term corresponding to the variance of the population prevalence p (Böhning et al., 1998). This variance is denoted by τ^2. In formula,

$$\mathrm{VAR}_{\text{het}}(\hat{p}_{\text{pool}}) = \hat{p}/N + \tau^2(n_1^2 + \ldots + n_k^2)/N^2 \tag{3}$$

Here, $\hat{p}$ is the overall prevalence (the weighted mean of the CL prevalences) and τ^2 the variance of the population prevalence. Obviously, the variance in Eq. (3) consists of two terms: the first one is due to the random variability within each cluster (it is the binomial variance approximated here by the Poisson variance) and a second term due to the heterogeneity between clusters (the variation of the CL prevalence parameter in the population). The term τ^2 can be obtained from the parameters of a mixing distribution as shown under Section 3.4. Eq. (3) demonstrates clearly that if population heterogeneity is ignored (τ^2>0), the variance of the prevalence estimator is underestimated by the term $\tau^2(n_1^2 + \ldots + n_k^2)/N^2$. Also, if there is population homogeneity (τ^2=0), both approaches and formulae coincide.

Approaches differ in the way τ^2 is estimated. According to the ICC approach outlined in the previous section, τ^2 is estimated in the parametric Beta distribution as $\hat{\rho}\hat{p}(1-\hat{p})$ (which is 0.0138 for our data set). In Section 3.4, a non-parametric approach for estimating iτ^2 is outlined.

Suppose for the moment, that a non-parametric estimator of τ^2 would be available (given in Section 3.4). Then, according to Eq. (3), the estimated variance that accounts for heterogeneity in our study data is $\mathrm{VAR}_{\text{het}}(\hat{p}_{\text{pool}})$=0.00076; the corresponding variance-inflation factor (established as the ratio of $\mathrm{VAR}_{\text{het}}$ and the Poisson variance for simple random sampling) is 2.07. We denote this as 'heterogeneity-approach'. Incorporation of heterogeneity leads to the confidence interval CI_{het}=[0.1246, 0.2327]. The different widths of naive, design-based and data-based (accounting for heterogeneity) confidence intervals are visualised in Fig. 2. The CIs based on naive (i.e. assumption of

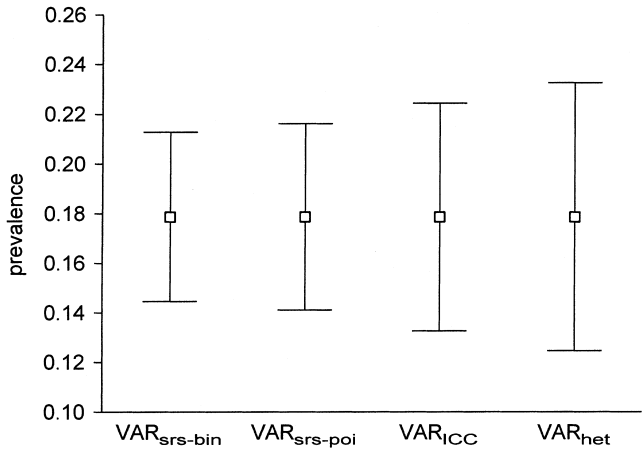

Fig. 2. Confidence intervals for the prevalence of bovine trypanosomosis in Mukono County, Uganda (data from June 1994, total sample size 487; 50 farms; mean farm size 9.7; intracluster correlation 0.0939) with 95% confidence intervals (CI, Normal approximation). The underlying variance was estimated under assumption of simple random sampling ($VAR_{srs\text{-}bin}$ for Binomial and $VAR_{srs\text{-}bin}$ for Poisson approximation), cluster sampling (VAR_{ICC}), and heterogeneity (VAR_{het}).

simple random sampling) variance estimators are smaller than those based on the design-based (i.e. accounting for intracluster correlation) variance. The CI based on the variance estimation that accounts for heterogeneity is even wider than that based on the intracluster correlation coefficient.

3.3. Comparison of the two variance inflation approaches

Two approaches for estimating the inflated variances need to be compared. For simplification, we used the Poisson approximation p/N of the Binomial variance $\hat{p}(1-\hat{p})/N$. We consider a cluster sampling with k clusters. According to Eq. (1), we estimate the inflation factor as $\hat{c} = 1 + \hat{\rho}(\bar{n} - 1)$ and the inflated variance becomes

$$(\hat{p}/N)(1 + \hat{\rho}(\bar{n} - 1)) = \hat{p}/N + \hat{p}\hat{\rho}(\bar{n} - 1)/N \approx \bar{p}/N + \tau^2/k$$

where we have used that for the Beta distribution, the variance $\tau^2 = \hat{\rho}\hat{p}(1-\hat{p}) \approx \hat{\rho}\hat{p}$ and the more severe assumption that all $n_i = \bar{n}$ for all i=1,..., k.

According to Eq. (3) the inflated variance becomes

$$\hat{p}/N + \tau^2(n_1^2 + \ldots + n_k^2)/N^2 \approx \hat{p}/N + \tau^2 k\bar{n}^2/(k\bar{n})^2 = \hat{p}/N + \tau^2/k$$

and both approaches coincide. Clearly, to achieve this result, the same sample size for the clusters had to be assumed. This might even introduce a more-severe bias than the differences due to using different estimators of τ^2. Indeed, a direct and simple argument shows that for all possible combinations of cluster sizes n_1, n_2,...,n_k

$$(n_1^2 + +n_k^2)/N^2 \geq 1/k$$

where the inequality is binding (becoming equality) if all cluster sizes coincide. This implies that $(1 + \hat{\rho}(\bar{n} - 1))$ underestimates the variance inflation factor – in case of unequal cluster sample sizes – potentially even drastically. Consequently, the heterogeneity-approach using Eq. (3) should be preferred since it is more realistic. If the cluster sizes are similar, the approaches will not be too different.

3.4. Diagnosis of heterogeneity

Frequently, the heterogeneity in the distribution of CL prevalences is so striking that simple graphical methods (such as frequency-distribution histograms and density traces) provide non-parametric density estimates of the involved subpopulations. From the density trace of CL prevalences of our study data (Fig. 1(B)), it is evident that not only one mode is present in the distribution but about three, one at zero, the second around 0.15, and the third around 0.35. Having more than one mode in the distribution of CL prevalences – as in our example – is what we denote as population heterogeneity in the context of prevalence estimation under cluster sampling. If there is evidence for population heterogeneity the problem remains how this heterogeneity can be estimated. It has been pointed out that the appropriate solution to this problem is a mixture model with unspecified mixing distribution giving weights w_1, w_2,..., w_m to the parameters (prevalences) $p_1, p_2, \ldots, p_m$ (Böhning et al., 1992). Heterogeneity can be interpreted as the distribution of a parameter in the (super-) population that consists of subpopulations with different mean parameter values. These subpopulations are having (prevalence) parameters $p_1, \ldots, p_m$ and the (super-) population is partitioned into these subpopulations according to the weights $w_1, w_2, \ldots, w_m$. Note that 'weight' refers here to the proportion of the identified subpopulation and is not to be confused with the sampling weights mentioned in the introduction. Conditionally, in subpopulation with index j, the number of cases y_i in cluster i is assumed to follow the Poisson distribution $\mathrm{Poi}(y_i, \lambda_{ij}) = \exp(-\lambda_{ij})\lambda_{ij}^{y_i}/y_i!$ with the parameter $\lambda_{ij} = p_j n_i$. The probability that the number of cases in herd i is given by $\Pr(Y_i = y_i) = \Sigma_j \Pr(Y_i = y_i |$ cluster $i \in$ subpopulation $j) \Pr($cluster $i \in$ subpopulation $j) = \Sigma_j \Pr(Y_i = y_i | p_j)\ w_j$. Now, the conditional probability is just $\Pr(Y_i = y_i | p_j) = \mathrm{Poi}(y_i, p_j n_i)$. Therefore, unconditionally, (i.e. not knowing the membership to the subpopulation), the following mixture model is valid for the number of cases y_i in herd i:

$$Pr(Y_i = y_i) = Poi(y_i, p_1 n_i)w_1 + Poi(y_i, p_2 n_i)w_2 + \ldots + Poi(y_i, p_m n_i)w_m$$

The parameters $p_1, \ldots, p_m$ and $w_1, \ldots, w_m$ as well as the number of subpopulations are estimated by a maximum-likelihood method using algorithms described by Böhning et al. (1992). In fact one considers the log-likelihood function $\Sigma_i \log[\mathrm{Poi}(y_i, p_1 n_i)w_1 + \ldots + \mathrm{Poi}(y_i, p_m n_i)w_m]$ as a function of $p_1, \ldots, p_m$, $w_1, \ldots, w_m$, m and uses optimisation procedures to find those values of the parameters p_j, w_j, $j = 1, \ldots, m$ and m itself, which maximise the log-likelihood.

For our study's data, the mixture analysis identified three subpopulation with the weights 0.1690, 0.4753, 0.3557 and the parameters 0.0, 0.1161, 0.3195, respectively. This means that about 17% of the clusters (herds) are infection-free, 48% have an infection prevalence of 12%, and 36% of the clusters show an infection prevalence

of 32%. From this heterogeneity distribution, mean and variances can easily be calculated leading to

$$\hat{p} = w_1 p_1 + w_2 p_2 + w_3 p_3$$

and

$$\hat{\tau}^2 = w_1(p_1 - \hat{p})^2 + w_2(p_2 - \hat{p})^2 + w_3(p_3 - \hat{p})^2 \quad (4)$$

In our case, we find $\hat{\tau}^2$=0.01428 – which is somewhat larger than the Beta-binomial based estimate of τ^2.

4. Discussion

The Beta-binomial distribution is considered a useful model of a discrete frequency distribution under cluster sampling in the presence of a positive intracluster correlation (i.e. extrabinomial variation) (e.g. Donald et al., 1994). Maximum-likelihood estimators are available to estimate the parameters of the Beta-binomial (i.e. p[overall prevalence] and Θ [index of aggregation]), from empirical data (Smith, 1983; Madden and Hughes, 1994). A two-parameter model, however, apparently imposes limitations to the structure of CL prevalences that can be appropriately modelled. We were interested in the estimation of the design-based variance of the binomial parameter p. Our data suggest that this variance – although accounting for extrabinomial variation – might still underestimate the data-based variance. This situation is due to heterogeneity in the distribution of CL prevalences as previously described by Böhning et al. (1998). The variance inflation based on heterogeneity can be expressed numerically if the parameters that describe heterogeneity (i.e. the number of subpopulations of clusters, their means and weights) are known or estimated. The diagnosis of heterogeneity – as well as the estimation of the involved parameters (number of subpopulations of clusters, the weights and mean prevalences if identified subpopulations) – is addressed by the concept of mixture-distribution analysis. Computer software that – besides the Binomial case – handles other distribution types as well (including Normal, Poisson,...) is available (Böhning et al., 1992) and has been used for diagnosis of heterogeneity in the context of seroepidemiology (Greiner et al., 1994, 1997).

The variance inflation due to heterogeneity should be distinguished from the inflation due to clustering in general – the latter being used frequently in an unspecific way. We have outlined above that the Beta-binomial (clustering) approach may not be sufficient in the presence of heterogeneity. The effect of variance inflation has been reviewed for a series of studies published in Preventive Veterinary Medicine – covering a wide range of epidemiological surveys (McDermott and Schukken, 1994). The cluster effect partially invalidated inferences from studies if the cluster sampling had not been accounted for in analysis. Additional variance inflation due to heterogeneity could be thought of producing the same trend of bias towards small p-values. We could not adjust our point estimate of the overall prevalence because information on the selection probabilities in

the sampling strata was missing. The resulting bias is probably mild since no correlation was found between cluster size and prevalence.

5. Conclusion

Cluster sampling may be perceived as an undesired but inevitable complexity in the analysis of veterinary epidemiological studies – e.g. if one is concerned with sample-size limitations. On the other hand, a certain form of cluster sampling is necessary to investigate the distribution of disease. Ignoring cluster sampling, thus, not only potentially invalidates study inferences but also leads to a failure to reveal the pattern of disease. This aspect may be even more important than getting proper variance estimates. A situation of distinct bi- or multimodal shapes of CL prevalences should be the starting point to investigate explanatory factors for different CL prevalence levels. Stratification of clusters according to such factors could potentially both enhance the statistical power of hypothesis testing and improve the understanding of the underlying biological background. The mixture distribution approach presented in this paper provides a suitable tool for the detection of heterogeneity and is, therefore, potentially of great practical significance.

With regard to the findings of this study, we suggest to consider the following eight points for the estimation of a prevalence under cluster sampling. (1) What kind of distribution of CL prevalences is expected? In case of highly contagious infections, a high intracluster correlation could be expected, where some clusters are having high prevalences and others are infection-free. On the other had, spontaneous disease outbreaks with low contagiousness would probably go along with a homogeneous distribution of CL prevalences. (2) A sampling strategy should be used that is suitable to verify the presumed kind of CL prevalence distribution. Generally, sampling should include clusters such as litters, flocks farms or herds. (3) The information of CL prevalences should be available (no aggregation of data without checking the distribution of CL prevalences). (4) Analysis of the data with descriptive and explorative methods. These methods include frequency-distribution graphs (kernel density estimates and histograms) of the CL prevalences. The relation of cluster size and prevalence may be analysed using linear regression. (5) Formal tests may be used to detect heterogeneity (i.e., extra-binomial variation or 'overdispersion'). A χ^2-test may be based on the (observed) calculated variance of the proportions divided by the expected value for the binomial distribution. Software is available for this purpose (Madden and Hughes, 1994). (6) In case of evidence for a heterogeneous distribution of CL prevalences (steps 4 and 5), the parameters of the mixed distribution could be found by mixture analysis. The computer-assisted mixture analysis (C.A.MAN, Böhning et al., 1992) provides maximum-likelihood estimates of the number of subpopulations (according to prevalence levels), their means (prevalences) and weights (proportions of the total population). (7) The variance of the overall prevalence that accounts for heterogeneity can be established using the results of step 6. We provide a closed form of this estimate Eq. (3). (8) The estimate of variance so established should be the basis for any inferences from the study data (construction of confidence intervals and statistical test).

Acknowledgements

We thank Prof. D. Mehlitz, Dr. R. Patzelt and all others involved in a BMZ funded Special Project (PN 94.7860.3-01.100) for the original data set which we used as example in our study. We also thank the Editor-in-chief and the two reviewers for their detailed and helpful comments. The research of the first author is under current support of the German Research Foundation.

Appendix

Eq. (1) provides the formula for the variance-inflation factor c which we have used for the ICC-approach of adjusting the simple random sampling variance. In the following, we outline the formal justification of this approach. First of all we note that the intracluster correlation addresses the within-cluster correlation. Therefore, we are concerned with a statistical model of autocorrelation in a single, hypothetical cluster. We call this cluster '*i*' from now on. Let y_i denote the number of cases out of n_i observations for the *i*th cluster with $i=1,\ldots,k$ and let X_{ij} denote an indicator variable for the *j*th observation from the *i*th cluster that takes the value 1 if the observation is a case and the value 0 otherwise, such that $y_i = \sum_{j=1}^{n_i} X_{ij}$.

Using the cluster-level prevalence $\Pr(X_{ij}=1)=p_i$, we can write the variance of X_{ij} as $\mathrm{VAR}(X_{ij})=p_i(1-p_i)$ and the expected value of the number of cases in the *i*th cluster as $E(Y_i)=n_i p_i$. We are interested in the components of the variance of Y_i

$$\begin{aligned}\mathrm{VAR}(Y_i) &= \mathrm{VAR}\left(\textstyle\sum_{j=1}^{n_i}(X_{ij})\right)\\ &= \textstyle\sum_{j=1}^{n_i}\mathrm{VAR}(X_{ij}) + \sum_{j=1}^{n_i}\sum_{a\neq j}\mathrm{Cov}(X_{ij},X_{ia})\\ &= \textstyle\sum_{j=1}^{n_i} P_i(1-p_i) + \sum_{j=1}^{n_i}\sum_{a\neq j}\mathrm{Cov}(X_{ij},X_{ia}).\end{aligned} \tag{5}$$

In case of independent observations X_{ij}, the covariance term Cov(.,.) becomes zero and the variance $\mathrm{VAR}(Y_i)=n_i p_i\ (1-p_i)$. this situation applies to a simple random sampling strategy. Supposed, however, there is autocorrelation (i.e. intracluster correlation) with $\mathrm{Cov}(.,.)\neq 0$. The definition of the intracluster correlation coefficient ρ can be simplified because $\mathrm{VAR}(X_{ij})=\mathrm{VAR}(X_{ia})=p_i(1-p_i)$

$$\begin{aligned}\rho &= \mathrm{Cov}(X_{ij},X_{ia})/[\mathrm{VAR}(X_{ij})\mathrm{VAR}(X_{ia})]^{0.5}\\ &= \mathrm{Cov}(X_{ij},X_{ia})/p_i(1-p_i)\end{aligned} \tag{6}$$

We resolve Eq. (6) for Cov(.,.) and insert the result into Eq. (5). Thus, in the presence of positive autocorrelation (infection leads to infection; $\rho>0$) the naive variance is inflated by $[1+(n_i-1)\ \rho]$

$$\begin{aligned}\mathrm{VAR}(Y_i) &= \textstyle\sum_{j=1}^{n_i} p_i(1-p_i) + \sum_{j=1}^{n_i}\sum_{a\neq j}\rho p_i(1-p_i)\\ &= n_i p_i(1-p_i) + n_i(n_i-1)\rho p_i(1-p_i)\\ &= n_i p_i(1-p_i)[1+(n_i-1)\rho].\end{aligned}$$

For the analysis of data under cluster sampling, Eq. (1) uses the mean cluster size $\bar{n}$. The exact factor that accounts for unequal sample sizes is $[1+(\Sigma_i n_i(n_i-1)/N)\rho]$ where the sum is from 1 to k. Eq. (2) provides an estimate of ρ based on the information out of k cluster.

References

Brier, S.S., 1980. Analysis of contingency tables under cluster sampling. Biometrika 67, 591–596.

Böhning, D., Schlattmann, P., Lindsay, B., 1992. Computer-assisted analysis of mixtures (CA.MAN): statistical algorithms. Biometrics 48, 283–303.

Böhning, D., Dietz, E., Schlattmann, P., 1998. Recent developments in computer assisted analysis of mixtures. Biometrics, 54(2).

Donald, A.W., Gardner, I.A., Wiggins, A.D., 1994. Cut-off points for aggregate herd testing in the presence of disease clustering and correlation of test errors. Prev. Vet. Med. 19, 167–187.

Donner, A., 1993. The comparison of proportions in the presence of litter effects. Prev. Vet. Med. 18, 17–26.

Fleiss, J.L., 1981. Statistical methods for rates and proportions, New York, Wiley, 227 pp.

Greiner, M., Franke, C.R., Böhning, D., Schlattmann, P., 1994. Construction of an intrinsic cut-off value for the sero-epidemiological study of *Trypanosoma evansi* infections in a canine population in Brazil: a new approach towards an unbiased estimation of prevalence. Acta Trop. 56, 97–109.

Greiner, M., Bhat, T.S., Patzelt, R.J., Kakaire, D., Schares, G., dietz, E., Böhning, D., Zessin, K.H., Mehlitz, D., 1997. Impact of biological factors on the interpretation of bovine trypanosomosis serology. Prev. Vet. Med. 30, 61–73.

Kairisto, V., 1995. Optimal bin widths for frequency histograms and ROC curves. Clin. Chem. 41, 766–767.

Kish, L., 1965. Survey sampling, Wiley, New York, 162 pp.

Madden, L.V., Hughes, G., 1994. BBD-Computer software for fitting the Beta-binomial distribution to disease incidence data. Plant. Dis. 78, 536–540.

Madden, L.V., Hughes, G., 1995. Plant disease incidence: distributions, heterogeneity, and temporal analysis. Annu. Rev. Phytopathol. 33, 529–564.

McDermott, J.J., Schukken, Y.H., 1994. A review of methods used for cluster effects in explanatory epidemiological studies of animal populations. Prev. Vet. Med. 18, 155–173.

McDermott, J.J., Schukken, Y.H., Shoukri, M.M., 1994. Study design and analytic methods for data collected from clusters of animals. Prev. Vet. Med. 18, 175–191.

Smith, D.M., 1983. Maximum-likelihood estimation of the parameters of the beta binomial distribution. Appl. Stat. 32, 192–204.

Stata Corp., 1997. Stata Statistical Software. College Station, TX: Stata Corporation. (version 5.0).