Statistical approach to measure the efficacy of anthelmintic treatment on horse farms

A. N. VIDYASHANKAR[1]*, R. M. KAPLAN[2] *and* S. CHAN[1]

[1] *Department of Statistical Science, Cornell University Ithaca, NY 14853-4201, USA*
[2] *Department of Infectious Diseases College of Veterinary Medicine, University of Georgia Athens, GA 30602, USA*

(*Received 30 April 2007; revised 29 June 2007; accepted 29 June 2007; first published online 23 August 2007*)

SUMMARY

Resistance to anthelmintics in gastrointestinal nematodes of livestock is a serious problem and appropriate methods are required to identify and quantify resistance. However, quantification and assessment of resistance depend on an accurate measure of treatment efficacy, and current methodologies fail to properly address the issue. The fecal egg count reduction test (FECRT) is the practical gold standard for measuring anthelmintic efficacy on farms, but these types of data are fraught with high variability that greatly impacts the accuracy of inference on efficacy. This paper develops a statistical model to measure, assess, and evaluate the efficacy of the anthelmintic treatment on horse farms as determined by FECRT. Novel robust bootstrap methods are developed to analyse the data and are compared to other suggested methods in the literature in terms of Type I error and power. The results demonstrate that the bootstrap methods have an optimal Type I error rate and high power to detect differences between the presumed and true efficacy without the need to know the true distribution of pre-treatment egg counts. Finally, data from multiple farms are studied and statistical models developed that take into account between-farm variability. Our analysis establishes that if inter-farm variability is not taken into account, misleading conclusions about resistance can be made.

Key words: efficacy, anthelmintic resistance, horse, beta-binomial model, logit-normal model, bootstrap methods, error rates, power.

INTRODUCTION

In recent years, anthelmintic resistance in gastrointestinal nematode parasites of livestock has emerged as an important problem worldwide. Multiple-drug-resistant parasites threaten small ruminant industries in many areas of the world, and resistance in parasites of horses and cattle is reaching alarming levels (Kaplan, 2004). The theoretical gold standard for diagnosing resistance to anthelmintics is achieved by counting the total number of killed worms and live worms following treatment; however, these data can be obtained only by sacrificing the animals, which is unrealistic on a farm. The practical gold standard is to measure changes in the number of eggs being produced by the parasites; these data can be obtained by measuring the number of eggs in a sample of feces. This procedure is called the fecal egg count reduction test (FECRT) and is the most common means of determining whether resistance is present on a farm (Kaplan, 2002). However, the fecal egg count (FEC) data are a surrogate measurement, which are subject to many sources of variability. Furthermore, the correlation between this surrogate measurement and the number of worms that are actually present in a horse is known to be weak (Lyons *et al*. 1983; Klei, 1986).

In a FECRT, fecal egg counts are compared in the same animals both before and after treatment, or between control and treated groups at some established time-point after treatment. However, there are many sources of animal-related and farm-related variability in FEC data that can impact the interpretation of results, especially when many different farms are being studied. Some of the most important sources of variability are: non-Gaussian overdispersed distribution of parasites in host animals, causing large differences in pre-treatment values between animals on the same farm; differences in parasite infection intensities between farms, causing large differences in pre-treatment values between farms; inherent variability in parasite egg numbers within the fecal output of an animal, which results in the collection of non-uniform samples (Warnick, 1992); variability in fecal egg counts resulting from the non-uniform distribution of eggs in solutions used for fecal egg count analysis; overall health and body condition of animals that can impact drug pharmacokinetics and pharmacodynamics; differences in age, breed, and sex of animals both on and between farms; differences in nutritional programs

* Corresponding author: Department of Statistical Science, Cornell University Ithaca, NY 14853-4201, USA. Tel: +607 255 3759. Fax: +607 255 9801. E-mail: anv4@cornell.edu

Parasitology (2007), **134**, 2027–2039. © 2007 Cambridge University Press
doi:10.1017/S003118200700340X Printed in the United Kingdom

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

between farms; spatial differences due to location of farms; and temporal differences resulting from non-uniform sampling times. These cause the FECRT data to be skewed and multi-modal. To overcome the problem of skewness, log and arcsine data transformations have been suggested (Fulford, 1994; Pook *et al*. 2002).

Many studies have investigated these issues in sheep and results have been used to develop methods for the experimental design and data analysis of FECRT (Coles *et al*. 1992). Although imperfect, these recommendations are generally accepted by parasitologists as being a useful standard. Unfortunately, virtually nothing has been done to investigate these issues in horses and, therefore, no accepted standards exist for either the design or data analysis of FECRT on horse farms (Kaplan, 2002). Furthermore, FECRT studies on horse farms are fraught with more severe problems in study design and analysis due to large numbers of horses with FEC of zero, and often only small numbers of horses available to test. Consequently, the study design used for FECRT in horses tends to differ from that used in sheep. In addition, major differences exist between sheep and horses in the biology of the host-parasite relationship. This implies that the metric used to evaluate treatment effect needs to be different for sheep and horses, which merits new statistical methodology.

Parasite burdens in animals are characterized by highly aggregated distributions within host populations (Crofton, 1971; Shaw and Dobson, 1995). This overdispersed pattern can be described by the negative binomial distribution (Shaw and Dobson, 1995). However, factors responsible for generating these observed patterns of parasite dispersion have not been well described, and although models have been developed, producing a tractable mathematical model for host-microparasite dynamics that allows for both the origins and effects of aggregation is a difficult technical problem (Anderson and Gordon, 1982; Grenfell *et al*. 1995). In parasitological investigations in animals, the issue of parasite distribution among hosts can have important effects on the interpretation of data. This is often addressed either by using geometric means to normalize the data set, or by using sufficiently large treatment groups that minimize the effect of aggregation (Fulford, 1994).

The approach for evaluating FECRT data in horses is to examine the arithmetic sample mean for percentage reduction, while some studies have used logarithmic and arcsine transformations before calculating the mean. An arbitrary percentage of either 90% or 95% reduction is often used to declare resistance (Bauer *et al*. 1986; Coles *et al*. 1992; Craven *et al*. 1998; Varady *et al*. 2000) while some studies have used a reduction of 80% for declaring resistance, with resistance suspected if the percentage reduction is between 80% and 90% (Woods *et al*. 1998; Kaplan *et al*. 2004).

Such approaches are based on the presumed efficacy of the drug rather than on the true efficacy of the drug at the time of treatment, which is unknown. However, since the correlation between the number of worms killed and the fecal egg count reduction (FECR) is weak and the number of horses is small, the accuracy of these arbitrary assignments of resistance to farms is unclear. Thus what is required is an extensive simulation study, which by design reflects the truth, so that we may begin to understand (1) the role of efficacy in understanding resistance, (2) statistical methods to evaluate efficacy, and (3) the role of variability in understanding efficacy and resistance.

In this paper, first we develop a theoretical framework for understanding resistance and efficacy that is free of distributional assumptions on the egg count distributions. Second we develop a novel bootstrap-based algorithm to assess efficacy, and we compare our methods to the existing approaches using extensive simulations. For this reason we develop several novel simulation models to test our methods. Finally, we show that these models can in turn be used to model FECRT data. To illustrate this, we provide an example using data from 2 horse farms that were part of a larger study on anthelmintic resistance using the FECRT.

Admittedly, the use of simulations to study resistance is not new (e.g. Torgerson *et al*. 2005; Morgan *et al*. 2005). However, the focus of the simulations and the methods used to evaluate resistance are new. In particular, we focus on hypothesis testing and the use of bootstrap methodology in this context.

MATERIALS AND METHODS

Modelling efficacy of anthelmintic treatment

FECRT data, as described above, are obtained by counting the number of eggs in a fecal sample from a horse. Let N denote the random variable describing the number of eggs in a fecal sample. Let $h(.)$ denote the probability distribution of N. That is,

$$P(N=k)=h(k). \tag{1}$$

Let $\lambda=E(N)$ denote the population mean of the egg counts in a fecal sample while $\sigma^2=Var(N)$ denotes the population variance of the number of egg counts in a fecal sample.

The effect of treatment is to kill worms and eliminate eggs. However, the FECRT data do not give information about the actual number of worms killed. Thus, if the treatment was efficacious (that is, the worms were killed) then the egg counts would be small, while if the treatment was not efficacious, the egg counts will be large relative to pre-treatment

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

counts. However, the converse may not be true. A low egg count in a sample may not necessarily mean the treatment is efficacious. The small number may be caused due to various factors affecting the egg count.

We say that a treatment efficacy exhibited through FECR is p $(0<p<1)$, if $100\times p\%$ of the eggs are eliminated. This does not mean that the true efficacy of the drug is p nor does it say that resistance is $(1-p)$. As mentioned in the introduction, several factors contribute to the efficacy of the treatment and hence it is not easy to diagnose resistance without understanding efficacy.

In this work, we assume that the true efficacy is an unknown parameter. We remark that even though the true efficacy is p, this does not mean that in every fecal sample $100\times p\%$ of eggs will be eliminated. The actual variability and the sample size could lead to variations around that number. It is often the case that in small sample size experiments with 'high' efficacy, there would be differences between the observed efficacy and the true efficacy. Thus without a statistical procedure, it is impossible to identify and evaluate the true efficacy of the treatment. The question of diagnosing resistance is related to hypothesis testing concerning efficacy but is more complicated.

The post-treatment egg count is modelled based on the pre-treatment egg count. Let N denote the random variable describing the pre-treatment egg count. Then the post-treatment egg count is modelled as a binomial distribution with the parameters pre-treatment egg count number and the efficacy, that is

$$\begin{aligned} &Y|N : Bin(N, 1-p)\\ &N : G(\lambda, \sigma^2), \end{aligned} \tag{2}$$

where G is an integer valued random variable with population mean λ and population variance σ^2. The above model is based on the assumption that given the pre-treatment egg count, the elimination process acts independently on the existing eggs. This is definitely a simplifying assumption but helps illustrate the concepts. The above modelling implies that the expected post-treatment egg count is $\lambda(1-p)$. The variance of the post-treatment egg count is

$$Var(Y)=\lambda p(1-p)+\sigma^2(1-p)^2. \tag{3}$$

Frequently, N is modelled to be a Poisson random variable with mean λ. In this Poisson case, the variance of the post-treatment egg count reduces to $\lambda p(1-p)+\lambda(1-p)^2$, since for the Poisson distribution mean = variance = λ.

Due to the aggregation phenomenon present in the FEC data, the negative binomial distribution (Cornell, 2005) has been suggested as an alternative to model the fecal egg counts. The negative binomial distribution has 2 parameters, namely, the mean λ and the negative binomial constant r. The probability distribution is given by

$$P_{r,\lambda}(N=k)=\frac{\Gamma(r+k)}{\Gamma(r)k!}\left(\frac{r}{r+\lambda}\right)^r\left(\frac{\lambda}{r+\lambda}\right)^k, \tag{4}$$

where Γ is the gamma function. The mean of the negative binomial distribution is λ and the variance is $\lambda\left(\frac{r+\lambda}{r}\right)$. Thus, in this case, the variance of the post-treatment egg count becomes $\lambda p(1-p)+\lambda\left(\frac{r+\lambda}{r}\right)(1-p)^2$.

Thus, the variance of the post-treatment egg count critically depends on the statistical model used for the pre-treatment egg count. However, the number of horses available to check the appropriateness of this modelling assumption is too small. Hence, if the assumed model for the pre-treatment egg count is incorrect, the resulting inference concerning the true efficacy of a drug and the resistance would be incorrect. Therefore, it is imperative to have a statistical methodology that is independent of the distributional assumptions on the egg count distributions.

Statistical issues for single farm data

Let us describe the data set from a single farm containing M horses with a positive pre-treatment egg count. Let N_i and Y_i denote the number of eggs in a fecal sample taken from the horse i before and after the anthelmintic treatment, respectively. Then $X_i=N_i-Y_i$ represents the change in the number of eggs. We will assume that the pair (N_i, Y_i) are independent and identically distributed random variables and that the effect of treatment on all the horses on the farm are the same. The meaning of this assumption is that the pair (pre-treatment egg count, post-treatment egg count) has the same statistical distribution for all horses in a given farm; and the outcome for a particular horse is statistically independent of the outcome from any other horse in the same farm. However, the realized values of Y_i will depend on N_i. Let p_i denote the efficacy of the treatment on horse i. Our assumption that the effect of treatment (i.e. proportion of worms killed) is the same for all horses on a farm implies that the probability of eliminating the worm eggs is also the same for all horses on the farm. This assumption allows us to set $p_i=p$ for all $1\leqslant i\leqslant M$. Much of the work in this paper deals with estimation and hypothesis testing concerning p.

To address the practical question of identifying and diagnosing resistance, one needs to provide a statistical estimate of p and test the hypothesis concerning p using the data $(N_1,Y_1),\ldots(N_M,Y_M)$. We estimate p using the formula

$$\hat{p}=\frac{1}{M}\sum_{i=1}^{M}\frac{X_i}{N_i}. \tag{5}$$

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

This is an unbiased estimate of p and is also a consistent estimator of p (Casella and Berger, 2001). The differences in the pre-treatment egg counts of horses are attributed to natural variation and hence the randomness of the pre-treatment egg counts yields unbiasedness of $\hat{p}$. Let us call the quantity $\frac{X_i}{N_i}$ the empirical efficacy from horse i. The above estimate of p, namely $\hat{p}$, is an average of the empirical efficacies calculated from each horse. The empirical efficacies from each horse take into account the differential number of worms in horses. If the horse i has a smaller number of worms, this will correspond to smaller N_i and this will lead to giving higher weight to the horse i. However, the variability from this horse will also be larger since $Var\left(\frac{X_i}{N_i}|N_i\right) = N_i^{-1}p(1-p)$. Thus the estimator of the true efficacy can be viewed as a weighted average of the change in egg counts in horses, the weights being $\frac{1}{N_i}$. Furthermore, the population variance of this estimator is given by $\frac{1}{M}\theta p(1-p)$, where θ is the population mean of the inverse of the pre-treatment egg counts. It is usually very difficult to calculate θ even for some well-known distributions. Our bootstrap methodology described in the following section describes a scheme that avoids this issue for a very general class of distributions. We reiterate that the properties of the estimator of p, namely $\hat{p}$, defined in (5) above, hold without any assumption concerning the distribution of the pre-treatment fecal egg count data.

Alternatively, one could consider the following unbiased estimator

$$\hat{p}_1 = \frac{\sum_{i=1}^{M} X_i}{\sum_{i=1}^{M} N_i}.$$

This estimator assumes that there is no variability between the horses, and does not allow the modelling of variability. The estimator $\hat{p}_1$ is the maximum likelihood estimator of p if there is no variability between the horses.

We now turn to hypothesis testing concerning p. Let H_0 and H_1 denote the null and the alternative hypotheses concerning p, respectively. The null hypothesis for p is a statement concerning the possible value of p. The alternative hypothesis, as the name indicates, states alternative values for p and is the research hypothesis. Typically, one tests if the null hypothesis can be rejected in favour of the alternative (research) hypothesis using the data. Testing of the hypothesis leads to acceptance or rejection of the null hypothesis which could be in error. In fact, there are two types of error that can arise and these are called Type I and Type II error. We now briefly discuss the terms Type I error and Type II error.

Type I error is caused by rejecting the null hypothesis H_0, when H_0 is true; while Type II error is caused by not rejecting H_0 when the alternative hypothesis H_1 is true. Let the null hypothesis concerning the percentage reduction be $H_0: p = p_0$. Let T_n denote the test statistic for testing H_0. The rejection region for testing H_0 is given by $[T_n > c]$ where c is so chosen that $P(T_n > c|H_0) \leqslant \alpha$, where $P(.|H_0)$ is the probability when the null hypothesis is true. Since c depends on α, we will write c_α to denote c. Furthermore, since the null hypothesis is true implies $p = p_0$, the Type 1 error rate can be expressed as $P_{p_0}(T_n > c_\alpha)$. The power of a test is then $P_p(T_n > c_\alpha)$, which we denote by $\beta_n(p)$.

The number p_0 is the presumed efficacy of the drug at the time of treatment. For instance, it is believed that ivermectin is 99·9% effective. In this case, one takes $p_0 = 0{\cdot}999$. In general, the value of p_0 can be obtained either from historical data or regulatory records.

Thus, when testing the hypothesis concerning p for a single farm, one is testing if the treatment is as efficacious as suggested in that farm using the data. A veterinary advisor will typically be interested in the answer to this question, which would help in making appropriate drug selection and drug treatment decisions.

Multiple farms

Frequently, a regulatory agency, a scientist or a veterinarian is interested in understanding if a particular treatment has certain levels of efficacy before suggesting the treatment for widespread use. In these situations, it is important to test the hypothesis concerning p using data from several farms. This will help in identifying the overall efficacy rate of the treatment. The data set in this case is a collection of single farm data from several farms. More formally, let R denote the number of farms and M_i denote the number of horses with positive egg count from farm i. Let N_{ij} and Y_{ij} denote the number of eggs in a fecal sample taken from the horse j before and after anthelmintic treatment, respectively. Let $X_{ij} = N_{ij} - Y_{ij}$. Then, X_{ij} represents the number of eggs eliminated by treating the j^{th} horse on the i^{th} farm. The effect of anthelmintic treatment on horses within the farm is the same but could differ between farms. We will denote by p_i the true efficacy of treatment on the i^{th} farm.

Statistical models for multiple farms

A parsimonious model for modelling the variability between farms is to assume that the efficacies for each farm p_i's are independent and identically distributed (i.i.d.) random variables taking values between 0 and 1. More precisely, let p_i, $1 \leqslant i \leqslant R$ denote independent random variables from a distribution $h(.)$. Conditionally on p_i, N_{ij}, we model Y_{ij} to be binomially distributed with parameters N_{ij} and p_i.

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Symbolically,

$$\begin{aligned} &Y_{ij}|(N_{ij}, p_i) : Bin(N_{ij}, 1-p_i) \\ &N_{ij} : G_i(\lambda, \sigma^2) \\ &p_i : h(.) \end{aligned} \tag{6}$$

where p_i's are independent random variables taking values on (0, 1) with distribution $h(.)$ and N_{ij} are i.i.d. distributed integer valued random variables from the distribution G_i. $h(.)$ is called the mixing distribution. Our methodology for assessing the overall efficacy does not require any distributional assumption on G_i. The model described above is the so-called generalized mixed model. We now specify 2 important special choices for $h(.)$, namely the beta distribution and the logit-normal distribution.

Beta-binomial model

In this model we assume that p_i, $1 \leqslant i \leqslant R$ are i.i.d. with a beta distribution with parameters δ_1 and δ_2; that is, the function $h(.)$ is the density function of a beta random variable with parameters δ_1 and δ_2. Under this assumption, (McCulloch and Searle, 2001)

$$E(p_i) = \frac{\delta_1}{\delta_1 + \delta_2} \tag{7}$$

while,

$$Var(p_i) = \frac{\delta_1 \delta_2}{(\delta_1 + \delta_2)^2(\delta_1 + \delta_2 + 1)}. \tag{8}$$

Our model induces correlations in the changes in the egg counts in the fecal samples from different horses from the same farm. When N_{ij} and N_{ik} are equal to 1, the correlation can be explicitly obtained using the formula

$$Corr(X_{ij}, X_{ik}) = \frac{1}{\delta_1 + \delta_2 + 1}, \ \textit{where } j \neq k. \tag{9}$$

In the case when they are not equal to 1, the correlation is given by a complicated mathematical expression involving N_{ij} and N_{ik}. The overall proportion reduction is now given by $E(p_i)$ and our hypotheses concern the values taken by the ratio $\frac{\delta_1}{\delta_1 + \delta_2}$.

Logit-normal model

Frequently, it is difficult to model p to take values between 0 and 1. For this reason, one models $\log\left(\frac{p_i}{1-p_i}\right)$ to have a normal distribution. The function $p_i \to \log\left(\frac{p_i}{1-p_i}\right)$ is called the logistic function and the resulting model is called the logit-normal model. Symbolically,

$$\begin{aligned} &X_{ij}|p_i : Bin(N_i, p_i) \\ &\log\frac{p_i}{1-p_i} : N(\mu, \sigma^2). \end{aligned} \tag{10}$$

This model does not have closed form expressions for the mean and the variance of p_i, but can be computed using numerical or Monte-Carlo algorithms (McCulloch and Searle, 2001). Standard software like SAS (Proc NLMIXED and PROC GLIMMIX in version 9.1) produce estimates for the mean and variances.

Proposed methods for assessing efficacy

As mentioned in the Introduction section, assessing the efficacy of treatment using FECRT data is challenging due to the fact that it exhibits patterns of aggregation, multi-modality and skewness. In this section, we propose 2 new methods for analysing FECRT data sets. We also describe the correct method to implement the arcsine transformation and the logarithmic transformation methods. Whereas the arcsine and logarithmic transformation methods require that assumptions be made about the nature of the pre-treatment egg count distribution, the bootstrap methods do not require any distributional assumptions.

Bootstrap Method 1 for a single farm

Using the notations from above, note that the pre-treatment data are $\{N_1, N_2, \ldots, N_M\}$ while the post-treatment data are $\{Y_1, Y_2, \ldots, Y_M\}$. The estimator of p as given in (5) is

$$\hat{p} = \frac{1}{M}\sum_{i=1}^{M}\frac{X_i}{N_i}.$$

We now describe our bootstrap algorithm to test the null hypothesis concerning p, namely test $H_0 : p = p$ (Efron and Tibshirani, 1993). Let B denote the number of bootstrap samples. In this algorithm, we keep the pre-treatment data fixed throughout the algorithm, namely, pre $= [N_1, N_2, \ldots, N_M]$.

1. Set k = 1.
2. Generate post-treatment bootstrap sample for the i^{th} horse by simulating post-treatment data from binomial distribution with parameters N_i and p_0. We repeat this process for all the horses yielding the k^{th} bootstrap sample of $post^*_k = \{Y^*_{1k}, Y^*_{2k}, \ldots, Y^*_{Mk}\}$. Define, for $1 \leqslant i \leqslant M$, $X^*_{ik} = N_i - Y^*_{ik}$.
3. Define $\hat{p}^*(k) = \frac{1}{M}\sum_{i=1}^{M}\frac{X^*_{ik}}{N_i}$.

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

4. Calculate the test statistic $T^*(k) = \frac{\sqrt{M}(\hat{p}^*(k) - p_0)}{\sqrt{\hat{p}^*(k)(1-\hat{p}^*(k))\gamma_M}}$, where $\gamma_M = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}$ (McCulloch and Searle, 2001).
5. Increment k by 1.
6. If the new value of k is less than B return to step 2. Else stop.

The distribution of the bootstrap samples $T^*(1),\ldots,T^*(B)$ is called the bootstrap distribution and can be used to approximate the distribution of the test statistic $T_n = \frac{\sqrt{M}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})\gamma_M}}$. The inference for p is now based on this bootstrap distribution. For instance, the p-value of the test can be obtained as the probability, calculated under the bootstrap distribution, that the test statistic takes values at least T_n. The confidence interval can similarly be obtained using the quantile points from the bootstrap distribution.

The above algorithm is repeated 5000 times to evaluate the Type I error and the power of the proposed methodologies.

Bootstrap Method 2 for a single farm

The main difference between the bootstrap Methods 1 and 2 is that in Method 2 we do not hold the pre-data fixed. This offers a degree of robustness to the statistical distribution of the pre-treatment egg count data. Instead, we obtain bootstrap samples for the pre-data and use these along with the bootstrap Method 1 to construct confidence interval for p. The actual steps of the algorithm are as follows:

1. Set k = 1.
2. Generate pre-treatment bootstrap samples by sampling from $\{N_1, N_2, \ldots N_M\}$ with equal probability. Let us denote the pre-treatment bootstrap sample by

$$pre_k^* = \{N_{1k}^*, N_{2k}^*, \ldots, N_{Mk}^*\}. \tag{11}$$

3. Now repeat the steps 2 through 6 of the bootstrap Method 1 where N_i is replaced by N_{ik}^*.

Confidence intervals and tests of hypotheses are constructed exactly as in the bootstrap Method 1 using the $\hat{p}^*(k)$ constructed in the present algorithm.

Bootstrap methods for multiple farms

In the case of multiple farms, one first obtains the efficacy of each farm and repeats the bootstrap methods 1 or 2 for each farm. Finally, an appropriate model is fit using the bootstrap estimates to obtain the overall efficacy rate and the confidence interval for the overall efficacy rate.

Other methods

In this section we describe other methods that have been used to test hypotheses concerning p and which we used to compare our bootstrap methods. The methods that we focus on include t-test, t-test after arcsine transformation and t-test after a log transformation. The value of $\hat{p}$ is obtained using the formula (5) for single farms. In the case of multiple farms, $\hat{p}$ can be obtained using the standard statistical software like SAS (Proc GLIMMIX, Proc NLMIXED).

The t-test

The standard t-statistic for testing $H_0: p = p_0$, is given by

$$t = \frac{\sqrt{M}(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})\gamma_M}} \tag{12}$$

where γ_M is defined as above. It is well known that, under the null hypothesis, t is approximately distributed as a t-distribution with $(M-1)$ degrees of freedom. Therefore, the rejection region is $|t| > t_{M-1,\frac{\alpha}{2}}$, where $t_{M-1,\frac{\alpha}{2}}$ is the critical value corresponding to the Type 1 error rate of α.

Arcsine transformation

The test statistic, to test $H_0: p = p_0$, corresponding to data transformed using arcsine transformation is

$$t = \frac{sin^{-1}\left(\sqrt{\hat{p}}\right) - sin^{-1}\left(\sqrt{p_0}\right)}{\sqrt{\frac{1}{4M^2}\left(\sum_{i=1}^{M}\frac{1}{N_i}\right)}}, \tag{13}$$

where t is approximately distributed as a t-distribution with $(M-1)$ degrees of freedom. The denominator of the above test statistic is actually a first order approximation to the variance of the numerator obtained using the delta method. Even though the denominator of the above expression does not involve p, the second and higher order terms involve p. Therefore, the rejection region is $|t| > t_{M-1,\frac{\alpha}{2}}$, where $t_{M-1,\frac{\alpha}{2}}$ is the critical value corresponding to the Type 1 error rate of α.

Log transformation

The test statistic, to test $H_0: p = p_0$, corresponding to data transformed using log transformation is

$$t = \frac{\sum_{i=1}^{M}(log(X_i) - log(N_i p_0))}{\sqrt{\sum_{i=1}^{M}\frac{\hat{q}}{N_i\hat{p}}}}, \tag{14}$$

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 1. Type I error rates for a single farm at varying levels of H_0 for five analysis methods examined

Distribution	p_0	*t*-test	Bootstrap1	Boostrap2	arcsin	log
Poisson	0·9	0·0388	0·0532	0·0516	0·0208	0·0418
$\lambda=12$	0·92	0·0446	0·0462	0·0494	0·0222	0·0452
	0·95	0·0616	0·0434	0·0456	0·0282	0·0614
	0·97	0·0880	0·0250	0·0240	0·0610	0·0888
	0·98	0·1518	0·0228	0·0202	0·1444	0·1518
Poisson	0·9	0·0192	0·0520	0·0520	0·0168	0·0190
$\lambda=100$	0·92	0·0206	0·0506	0·0512	0·0166	0·0208
	0·95	0·0276	0·0578	0·0576	0·0216	0·0282
	0·97	0·0268	0·0504	0·0490	0·0208	0·0270
	0·98	0·0324	0·0540	0·0510	0·0226	0·0322
Poisson	0·9	0·0186	0·0500	0·0490	0·0190	0·0188
$\lambda=500$	0·92	0·0182	0·0528	0·0518	0·0178	0·0182
	0·95	0·0188	0·0484	0·0472	0·0170	0·0186
	0·97	0·0152	0·0482	0·0492	0·0148	0·0150
	0·98	0·0174	0·0508	0·0500	0·0176	0·0172
Negative	0·9	0·0160	0·0464	0·0466	0·0132	0·0158
Binomial	0·92	0·0236	0·0464	0·0484	0·0202	0·0240
$\lambda=r=60$	0·95	0·0256	0·0530	0·0518	0·0200	0·0260
	0·98	0·0432	0·0494	0·0512	0·0190	0·0430
	0·99	0·0650	0·0516	0·0568	0·0364	0·0648
Negative	0·9	0·0198	0·0526	0·0524	0·0188	0·0202
Binomial	0·92	0·0180	0·0492	0·0486	0·0180	0·0178
$\lambda=r=250$	0·95	0·0196	0·0522	0·0528	0·0184	0·0196
	0·98	0·0208	0·0510	0·0526	0·0192	0·0206
	0·99	0·0326	0·0522	0·0542	0·0212	0·0328
Negative	0·9	0·0174	0·0510	0·0496	0·0180	0·0174
Binomial	0·92	0·0188	0·0544	0·0532	0·0178	0·0192
$\lambda=r=500$	0·95	0·0188	0·0516	0·0532	0·0184	0·0184
	0·98	0·0214	0·0512	0·0508	0·0198	0·0214
	0·99	0·0232	0·0518	0·0508	0·0208	0·0232
Negative	0·9	0·0176	0·0532	0·0544	0·0172	0·0174
Binomial	0·92	0·0156	0·0516	0·0520	0·0160	0·0156
$\lambda=r=1000$	0·95	0·0236	0·0532	0·0518	0·0224	0·0232
	0·98	0·0190	0·0468	0·0486	0·0166	0·0188
	0·99	0·0236	0·0574	0·0552	0·0206	0·0236

Column 1 represents the distribution with which the pre-data are generated.
Column 2 represents null hypothesis values of p.

where $\hat{q}=1-\hat{p}$ and t is approximately distributed as a t-distribution with $(M-1)$ degrees of freedom. Therefore, the rejection region is $|t|>t_{M-1,\frac{\alpha}{2}}$, where $t_{M-1,\frac{\alpha}{2}}$ is the critical value corresponding to the Type 1 error rate of α.

Throughout this paper we will choose $\alpha=0{\cdot}05$ and $B=2000$.

Simulation models and methods for single and multiple farms

In this section, we describe our basic setup for performing simulations. We adopted a setup that reflects the reality as closely as possible. In the discussions below, the means (λ) used for the simulations correspond to the observed raw egg counts. The number of eggs per gram (EPG) of feces is then obtained using the formula *sensitivity* $\times\lambda$. In our lab sensitivity = 5.

Single farm. In several of the resistance studies on horse farms investigated by the authors, data with a very small to moderate pre-treatment fecal egg count was common. For this reason, the pre-treatment data were generated from a Poisson distribution with mean λ, where $\lambda=2$ and negative binomial distribution with $\lambda=60$ and $r=60$. These numbers correspond to the 60 and 300 EPG. We also performed simulations in farms with large to very large pre-treatment egg counts. These included data from Poisson distribution with $\lambda=100$ and $\lambda=500$ and negative binomial distribution with $\lambda=250$ ($r=250$), $\lambda=500$ ($r=500$), and $\lambda=1000$ ($r=1000$). These correspond to between 500 and 5000 EPG of feces. Since the number of horses per farm was small, we chose 8 horses per farm. This yields pre-treatment data $N_1, N_2, \ldots, N_8$. Since the Type I error rate is defined to be the probability of rejecting H_0 when $H_0: p=p_0$ is true, the post-treatment data

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 2. Type I error rates for both single farm and multiple farms at varying levels of H_0 for five analysis methods examined

	p_0	*t*-test	Bootstrap1	Bootstrap2	arcsin	log
Single farm	0·80	0·0260	0·0474	0·0472	0·0172	0·0278
	0·85	0·0320	0·0518	0·0540	0·0212	0·0356
	0·90	0·0388	0·0532	0·0516	0·0208	0·0418
	0·95	0·0616	0·0434	0·0456	0·0282	0·0614
	0·98	0·1518	0·0228	0·0202	0·1444	0·1518
Multiple farms	0·80	0·0680	0·0758	0·0760	0·0694	0·1504
Beta-binomial model	0·85	0·0794	0·0832	0·0840	0·0788	0·1300
Variance between farms = 0·0004	0·90	0·0880	0·0972	0·0932	0·0888	0·1286
	0·95	0·1356	0·1366	0·1376	0·1310	0·1584
	0·98	0·2492	0·2412	0·2448	0·2424	0·2704
Multiple farms	0·80	0·0658	0·0678	0·0690	0·0628	0·1336
Logistic normal model	0·85	0·0614	0·0684	0·0670	0·0604	0·1146
Variance between farms = 0·0004	0·90	0·0770	0·0816	0·0796	0·0746	0·1142
	0·95	0·1166	0·1194	0·1202	0·1136	0·1378
	0·98	0·1544	0·1472	0·1450	0·1414	0·1676
Multiple farms	0·80	0·2640	0·2722	0·2718	0·2602	0·3466
Beta-binomial model	0·85	0·2998	0·3040	0·3044	0·2968	0·3600
Variance between farms = 0·004	0·90	0·3654	0·3738	0·3736	0·3652	0·4188
	0·95	0·5046	0·5022	0·5016	0·4966	0·5364
	0·98	0·6946	0·6926	0·6930	0·6916	0·7236
Multiple farms	0·80	0·2334	0·2448	0·2448	0·2324	0·3166
Logistic normal model	0·85	0·2658	0·2706	0·2752	0·2634	0·3234
Variance between farms = 0·004	0·90	0·3096	0·3142	0·3134	0·3042	0·3590
	0·95	0·4324	0·4248	0·4292	0·4256	0·4610
	0·98	0·6326	0·6178	0·6166	0·6236	0·6578

Column 2 represents null hypothesis values of p.
Pre-data are generated from the Poisson distribution with $\lambda = 12$.

were generated, for the i^{th} horse, using a binomial distribution with parameters (N_i, p_0). Based on the pre-treatment and post-treatment data, bootstrap confidence intervals were constructed using $B =$ 2000 bootstrap samples. The Type I error rate was calculated to be the percentage of times the null hypothesis was incorrectly rejected amongst the several simulated data sets. All our results are based on 5000 simulations. Type I error rate was evaluated at $p_0 = 0{\cdot}90$, 0·92, 0·95, 0·97, and 0·98 for the Poisson distribution and $p_0 = 0{\cdot}9$, 0·92, 0·95, 0·98, and 0·99 for the negative binomial distribution.

Multiple farms. In this case our simulation results are based on 10 farms and 8 horses per farm. The pre-treatment data for each horse in a farm are simulated from a Poisson distribution with $\lambda = 12$. The post-treatment data are simulated using the beta-binomial model and the logit-normal, respectively. The Type I error rate was calculated to be the percentage of times the null hypothesis was incorrectly rejected amongst the simulated data sets. All the results were based on 5000 simulations and the Type I error was evaluated at $p_0 = 0{\cdot}80$, 0·85, 0·90, 0·95, and 0·98. These p_0 values represent the population averages and the actual efficiency varied between farms around p_0.

Simulated power. The power of the proposed bootstrap methods and other methods can be obtained using the simulations. To calculate the simulated power, the post-treatment data are simulated using the probability in the alternative hypothesis and the bootstrap power is defined to be the proportion of times the false null hypothesis is rejected amongst the simulated data sets.

RESULTS

Results for Type I error rate

Table 1 presents Type I error rate results for the proposed bootstrap methods and other methods for data from a single farm for various values of the pre-treatment egg count mean. From the table it is clear that the Type I error rates are close to the nominal 5% level for the bootstrap method while for all other methods the Type I error rates are less than the nominal 5% level, as long as the mean raw pre-treatment egg count is greater than 60. If the pre-treatment egg count is small, then bootstrap methods still yield optimal Type I error rates if the true efficacy is less than or equal to 95%. However, for larger efficacy rates while bootstrap methods yield lower Type I error rates, other methods yield higher Type I error rates. Results in Table 2 show that if the

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 3. Power function for a single farm at varying levels of the true efficacy rate p

H_0	p	t-test	Bootstrap1	Bootstrap2	arcsin	log
$p_0=0{\cdot}95$	0·89	0·9846	0·9978	0·9970	0·9906	0·9848
	0·9	0·9272	0·9848	0·9858	0·9524	0·9288
	0·92	0·5028	0·7690	0·7696	0·6052	0·5092
	0·93	0·2114	0·4754	0·4780	0·2928	0·2162
	0·94	0·0404	0·1700	0·1698	0·0734	0·0434
$p_0=0{\cdot}97$	0·9	0·9998	1·0000	1·0000	1·0000	0·9998
	0·92	0·9788	0·9974	0·9970	0·9912	0·9796
	0·93	0·9014	0·9824	0·9822	0·9484	0·9014
	0·94	0·6650	0·8898	0·8890	0·7696	0·6694
	0·95	0·2962	0·6236	0·6228	0·4210	0·2994
	0·96	0·0576	0·2464	0·2478	0·1128	0·0594
$p_0=0{\cdot}99$	0·9	1·0000	1·0000	1·0000	1·0000	1·0000
	0·92	1·0000	1·0000	1·0000	1·0000	1·0000
	0·95	0·9884	0·9998	0·9996	0·9978	0·9890
	0·96	0·9042	0·9922	0·9914	0·9690	0·9078
	0·97	0·5816	0·9012	0·8994	0·7740	0·5850
	0·98	0·1050	0·4824	0·4794	0·2620	0·1068

Pre-data are generated using negative binomial distribution with $\lambda=60$ and $r=60$.

variability is not taken into account in the analysis, Type I error rates increase for all the methods.

Power results

Table 3 presents the power analysis for a single farm data when the mean of the pre-treatment raw egg count distribution is 60 or equivalently 300 eggs per gram (EPG) of feces. The results clearly show that both the bootstrap methods have substantial power to detect changes close to the null hypothesis as compared to the other methods. Tables 4 and 5 give the simulated power results for data from multiple farms with different distributions for variability. Again these results demonstrate that the bootstrap method has 'reasonable' power to detect differences from the null hypothesis.

Figure 1 shows that the distribution of the change in egg count has an approximately similar shape for various efficacy levels and for various pre-treatment egg count mean levels. This shows why detecting even 5% change in efficacy is such an arduous task in these data.

Real data example

In this section we analysed a real data set that was collected as part of a study on anthelmintic efficacy across various farms in the southeastern United States (Kaplan *et al*. 2004). For this illustration, we focused on the farms in the state of Louisiana. Nine farms were included in the Louisiana study and horses on each farm were assigned randomly to one of several anthelmintic treatments. One of the treatments was ivermectin. We analysed the data for all the 9 farms. All of the farms except ASHU and EHS had 100% reduction. Hence, p-values and confidence intervals are provided for only ASHU and EHS. On the farms with 100% reduction, since there is no variation in the data, the lower and upper confidence limits coincide and equal 1. The 95% confidence intervals and p-values for test of $H_0: p=0{\cdot}98$ are given in Table 6.

The bootstrap confidence interval for percentage reduction for ASHU farm is (0·962, 0·971) indicating, perhaps, the beginning stages of resistance, while that of EHS is (0·995, 1·0) indicating that the efficacy was very high. However, for the ASHU farm, the confidence interval that takes into account the variability between farms was determined to be (0·992, 0·999) yielding that there were no initial stages of resistance. To validate these results, the experiment was repeated at the ASHU farm for ivermectin using an increased number of horses and no resistance was detected.

DISCUSSION

The work presented in this paper addresses 4 fundamental issues. First, development of a statistical model for understanding efficacy. Second, methods to analyse efficacy on farms with small pre-treatment egg counts with small number of horses. Third, the role and desirability of data transformations; and fourth, the impact of variability between farms on the accurate measurement of efficacy.

Our theoretical framework assumes a conditional independence model for elimination of eggs and clarifies differences between resistance and efficacy. A population model is assumed and all statistical analyses are made relative to the assumed population model. The role of variability in the initial egg count

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 4. Power function for multiple farms with data generated using the beta-binomial model at varying levels of the efficacy rate p

Variance	H_0	p	t-test	Bootstrap1	Bootstrap2	arcsin	log
0·0004	$p_0=0·9$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	0·9834	0·9864	0·9870	0·9856	0·9956
		0·90	0·0880	0·0972	0·0932	0·0888	0·1286
		0·95	0·9978	0·9976	0·9974	0·9976	0·9942
		0·98	1·0000	1·0000	1·0000	1·0000	1·0000
	$p_0=0·95$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	1·0000	1·0000	1·0000	1·0000	1·0000
		0·90	0·9972	0·9984	0·9982	0·9978	0·9994
		0·95	0·1356	0·1366	0·1376	0·1310	0·1584
		0·98	0·9688	0·9610	0·9616	0·9620	0·9538
0·004	$p_0=0·9$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	0·9998	1·0000	1·0000	1·0000	1·0000
		0·85	0·8854	0·9006	0·9030	0·8938	0·9428
		0·90	0·3654	0·3738	0·3736	0·3652	0·4188
		0·95	0·9146	0·9104	0·9102	0·9098	0·8788
		0·98	0·9870	0·9858	0·9860	0·9860	0·9722
	$p_0=0·95$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	1·0000	1·0000	1·0000	1·0000	1·0000
		0·90	0·9448	0·9572	0·9584	0·9532	0·9688
		0·95	0·5046	0·5022	0·5016	0·4966	0·5364
		0·98	0·8682	0·8616	0·8618	0·8638	0·8690

Shaded row represents the Type I error rate.
Column 1 represents the variance between farms.
Pre-data are generated from the Poisson distribution with $\lambda=12$.

is brought out and its impact on assessing efficacy is studied.

FECRT data exhibit various peculiarities that make analysis particularly challenging and, for a number of reasons described previously, these difficulties are magnified in horses. Bootstrap methodology is a theoretically sound technique for performing inference with minimal assumptions on the statistical distribution of the data. The method is especially applicable to small data sets, namely data sets with small pre-treatment egg counts and few horses per farm, since one can approximate the sampling distribution of the test statistic by simulating a large number of bootstrap samples. Furthermore, the methodology is easily adaptable and implementable in complex problems involving complicated models since, unlike the other methods presented, bootstrap methods do not require calculation of the variance of the statistic as was shown in the previous section. All that is needed is to generate data from the specified model, and then calculation of the test statistic.

When studying resistance on a single farm, our results convincingly showed that bootstrap methods perform optimally both in terms of Type I error and the power, even when the pre-treatment egg counts are small. On farms with large pre-treatment egg counts, it is believed that t-tests would work well, since the t-distribution is 'well' approximated by the normal distribution. The statistical reason behind this belief is the so-called central limit theorem (Billingsley, 1995), which states that as the sample size n increases without bound, the difference in the probabilities calculated using the t-distribution with n degrees of freedom and the normal distribution decreases to 0. However, how fast the difference goes to 0 depends on the values of p_0. This phenomenon is well explained by our results in Tables 1 and 3.

The problem of assessing efficacy of treatment on farms with small egg counts is very difficult. Even on farms with mean egg counts of 100 or 200, it is hard to detect even a 5% drop in efficacy with 8 horses. This leads to a substantial difference in the presumed efficacy and the true efficacy of the treatment. A practical consequence of this effect is that resistance, if properly defined, will go undetected in many instances. The variability in the pre-treatment egg count compounds the problem and in some cases even a 10% drop in treatment efficacy can go undetected. Figure 1 reiterates this phenomenon.

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 5. Power function for multiple farms with data generated using the logistic normal model at varying levels of the efficacy rate p

Variance	H_0	p	t-test	Bootstrap1	Bootstrap2	arcsin	log
0·0004	$p_0=0·9$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	0·9864	0·9908	0·9904	0·9890	0·9968
		0·90	0·0770	0·0816	0·0796	0·0746	0·1142
		0·95	0·9992	0·9988	0·9988	0·9992	0·9978
		0·98	1·0000	1·0000	1·0000	1·0000	1·0000
	$p_0=0·95$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	1·0000	1·0000	1·0000	1·0000	1·0000
		0·90	0·9982	0·9990	0·9992	0·9990	0·9992
		0·95	0·1166	0·1194	0·1202	0·1136	0·1378
		0·98	0·9892	0·9854	0·9852	0·9866	0·9834
0·004	$p_0=0·9$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	0·9044	0·9178	0·9162	0·9116	0·9530
		0·90	0·3096	0·3142	0·3134	0·3042	0·3590
		0·95	0·9446	0·9410	0·9392	0·9400	0·9126
		0·98	0·9868	0·9864	0·9862	0·9862	0·9804
	$p_0=0·95$	0·70	1·0000	1·0000	1·0000	1·0000	1·0000
		0·75	1·0000	1·0000	1·0000	1·0000	1·0000
		0·80	1·0000	1·0000	1·0000	1·0000	1·0000
		0·85	1·0000	1·0000	1·0000	1·0000	1·0000
		0·90	0·9616	0·9700	0·9700	0·9678	0·9780
		0·95	0·4324	0·4248	0·4292	0·4256	0·4610
		0·98	0·8928	0·8870	0·8872	0·8876	0·8992

Shaded row represents the Type I error rate.
Column 1 represents the variance between farms.
Pre-data are generated from the Poisson distribution with $\lambda=12$.

When performing FECRT in horses, arbitrary minimum EPG cutoffs are typically used for accepting horses into the study. The logic used is that higher EPG will yield a more accurate measurement of efficacy. However, results of simulations challenge the validity of this logic. These simulation data demonstrate that reductions in fecal egg counts are inherently highly variable, causing difficulty in assessing the true efficacy. We can see that this phenomenon changes little as the EPG changes from 60 to 5000. The practical significance of this finding is that it is very difficult to distinguish a true egg count reduction of 90% from that of 95% when testing only small numbers of horses. Consequently, our results suggest that it is preferable to include as many horses as possible in a FECRT, even those with low EPG, and to use a more sensitive assay for measuring EPG.

In the context of multiple farms, our results clearly show that the variability between farms will have to be taken into account to detect efficacy. Not accounting for the variability usually leads to increased Type I error. The practical significance of this phenomenon is a more frequent, albeit incorrect, diagnosis of resistance leading to an overestimation of resistance prevalence.

The models introduced in this paper, namely the beta-binomial model and the logit-normal model, can be used to model variability between farms. With the help of these models one can estimate the true efficacy and perform hypothesis tests along the methods presented in the Materials and Methods Section. As mentioned previously, when performing hypothesis tests, the bootstrap method is significantly easier since it only requires simulating data according to the fitted model, unlike other methods which require complicated calculations to determine the variance.

Finally, using our theoretical framework it is possible to introduce a notion of resistance based on the estimate of efficacy and the presumed efficacy. In our limited simulation study, we see that large number of horses per farm are required before making unequivocal statements concerning resistance. We are currently addressing these and other pertinent issues using a number of different statistical approaches with the goal of improving our understanding of efficacy, and our understanding of

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

Table 6. Analysis of Louisiana farms for horse data

Farm		*t*-test	Bootstrap1	Bootstrap2	arcsin	log
ASHU	point estimate	0·96654	0·96654	0·96654	0·96654	0·96654
	lower CI	0·96193	0·96234	0·96249	0·96178	0·95861
	upper CI	0·97114	0·97079	0·97071	0·97099	0·96778
	p-value	0·000035	0·000000	0·000000	0·000011	0·000004
EHS	point estimate	0·99772	0·99772	0·99772	0·99772	0·99772
	lower CI	0·99400	0·99493	0·99424	0·99250	0·99399
	upper CI	1·00140	1·00020	1·00020	0·99992	1·00140
	p-value	0·000064	0·000000	0·000000	0·001580	0·000062

Treatment is ivermectin.
p-value is calculated under the null hypothesis H_0: $p = 0{\cdot}98$.

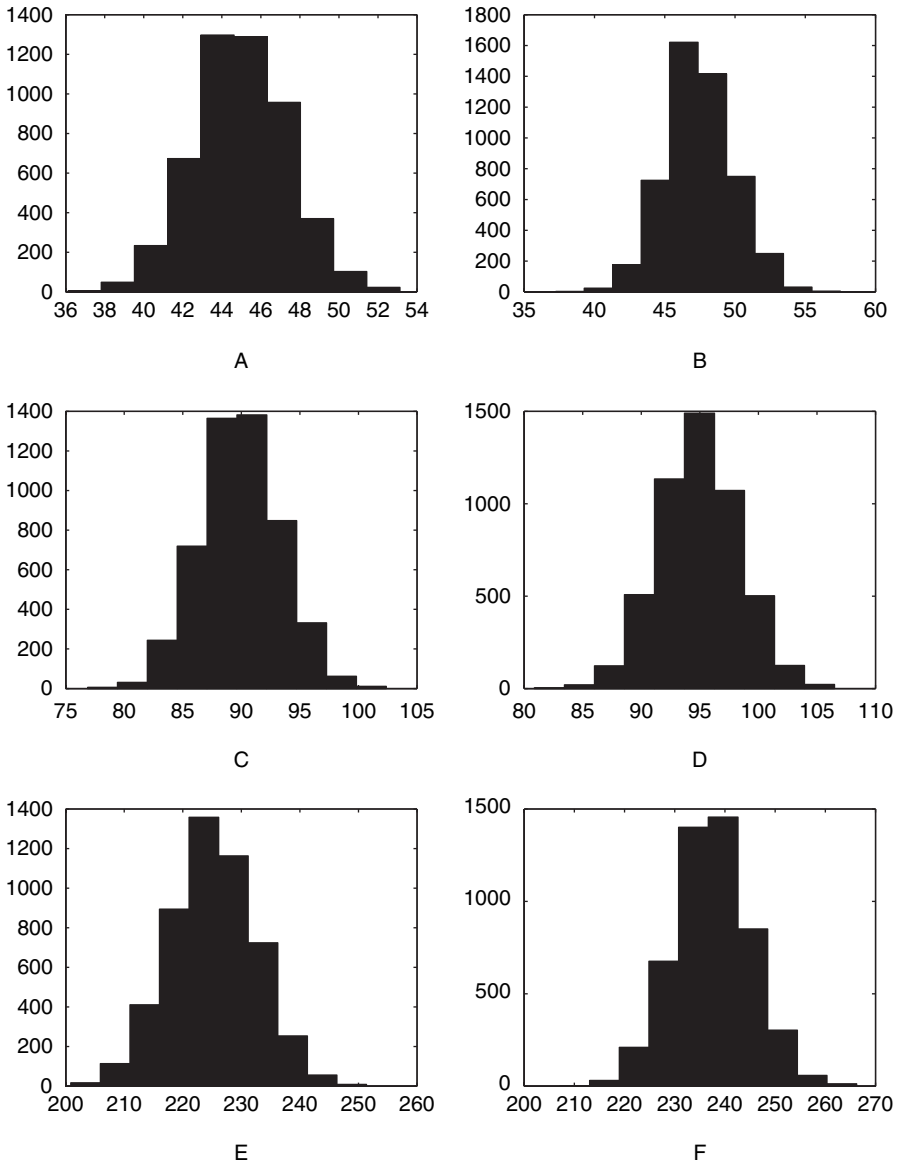

Fig. 1. Histograms of the simulated distribution of pre-treatment egg counts minus post-treatment egg counts. The pre-treatment egg counts were generated using various distributions, under the following true efficacy rates: (A) Poisson ($\lambda = 50$), $p = 0{\cdot}9$; (B) Poisson ($\lambda = 50$), $p = 0{\cdot}95$; (C) Poisson ($\lambda = 100$), $p = 0{\cdot}9$; (D) Poisson ($\lambda = 100$), $p = 0{\cdot}95$; (E) Negative Binomial ($\lambda = 250$, $r = 250$), $p = 0{\cdot}9$; (F) Negative Binomial ($\lambda = 250$, $r = 250$), $p = 0{\cdot}95$.

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press

how to make an accurate diagnosis of resistance on the basis of fecal egg count data.

REFERENCES

Anderson, R. M. and Gordon, D. M. (1982). Processes influencing the distribution of parasite numbers within host populations with special emphasis on parasite-induced host mortalities. *Parasitology* **85**, 373–398.

Bauer, C., Merkt, J., Janke-Grimm, G. and Burger, H. (1986). Prevalence and control of benzimidazole-resistant small strongyles on German thoroughbred studs. *Veterinary Parasitology* **21**, 189–203.

Billingsley, P. (1995). *Probability and Measure*, 3rd Edn. John Wiley & Sons, Inc, New York.

Casella, G. and Berger, R. L. (2001). *Statistical Inference*. Duxbury, New York.

Coles, G. C., Bauer, C., Borgsteede, F. H. M., Geerts, S., Klei, T. R., Taylor, M. A. and Waller, P. J. (1992). World Association for the Advancement of Veterinary Parasitology (W.A.A.V.P.) methods for the detection of anthelmintic resistance in nematodes of veterinary importance. *Veterinary Parasitology* **44**, 35–44.

Cornell, S. (2005). Modelling nematode populations: 20 years of progress. *Trends in Parasitology* **21**, 542–545.

Craven, J., Bjorn, H., Henriksen, S. A., Nansen, P., Larsen, M. and Lendal, S. (1998). Survey of anthelmintic resistance on Danish horse farms, using 5 different methods of calculating fecal egg count reduction. *Equine Veterinary Journal* **30**, 289–293.

Crofton, H. D. (1971). A quantitative approach to parasitism. *Parasitology* **62**, 179–193.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Fulford, A. J. C. (1994). Dispersion and bias: Can we trust geometric means? *Parasitology Today* **10**, 446–448.

Grenfell, B. T., Wilson, K., Isham, V. S., Boyd, H. E. G. and Dietz, K. (1995). Modelling patterns of parasite aggregation in natural populations: Trichostrongylid nematode-ruminant interactions as a case study. *Parasitology* **111** (Suppl.), S135–S151.

Kaplan, R. M. (2002). Anthelmintic resistance in nematodes of horses. *Veterinary Research* **33**, 491–507.

Kaplan, R. M. (2004). Drug resistance in nematodes of veterinary importance: a status report. *Trends in Parasitology* **20**, 477–481.

Kaplan, R. M., Klei, T. R., Lyons, E. T., Lester, G. D., French, D. D., Tolliver, S. C., Courtney, C. H., Vidyashankar, A. N. and Zhao, Y. (2004). Prevalence of anthelmintic resistant cyathostomes on horse farms. *Journal of the American Veterinary Medical Association* **225**, 903–910.

Klei, T. R. (1986). Laboratory diagnosis. In *Veterinary Clinics of North America: Equine Practice* (ed. Herd, R. P.), pp. 381–393. W.B. Saunders, Philadelphia, USA.

Lyons, E. T., Tolliber, S. C. and Drudge, J. H. (1983). Critical tests in equids with fenbendazole alone or combined with piperazine: particular reference to activity on benzimidazole-resistant small strongyles. *Veterinary Parasitology* **12**, 91–98.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Interscience, New York.

Morgan, E. R., Cavill, L., Curry, G. E., Wood, R. M. and Mitchell, E. S. E. (2005). Effects of aggregation and sample size on composite fecal egg counts in sheep. *Veterinary Parasitology* **131**, 79–87.

Pook, J. F., Power, M. L., Sangster, N. C., Hodgson, J. L. and Hodgson, D. R. (2002). Evaluation of tests for anthelmintic resistance in cyathostomes. *Veterinary Parasitology* **106**, 331–343.

SAS Publishing. (2004). *SAS/STAT Users Guide, Version 9.1*. SAS Institute Inc, Cary, NC, USA.

Shaw, D. J. and Dobson, A. P. (1995). Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology* **111** (Suppl.), S111–S133.

Torgerson, P. R., Schnyder, M. and Hertzberg, H. (2005). Detection of anthelmintic resistance: a comparison of mathematical techniques. *Veterinary Parasitology* **128**, 291–298.

Varady, M., Konigova, A. and Corba, J. (2000). Benzimidazole resistance in equine cyathostomes in Slovakia. *Veterinary Parasitology* **94**, 67–74.

Warnick, L. (1992). Daily variability of equine fecal strongyle egg counts. *The Cornell Veterinarian* **82**, 453–463.

Woods, T. F., Lane, T. J., Zeng, Q. Y. and Courtney, C. H. (1998). Anthelmintic resistance on horse farms in north central Florida. *Equine Practice* **20**, 14–17.

https://doi.org/10.1017/S003118200700340X Published online by Cambridge University Press