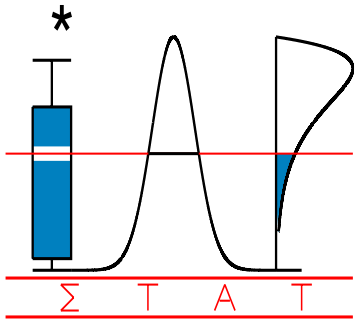
T E C H N I C A L

R E P O R T

0412

UNDERSTANDING HETEROGENEITY IN MIXED, GENERALIZED MIXED AND FRAILTY MODELS

Luc DUCHATEAU and Paul JANSSEN



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Understanding Heterogeneity in Mixed, Generalized Mixed and Frailty Models

Luc Duchateau

Department of Physiology, Biochemistry and Biometrics, Faculty of Veterinary Medicine, Ghent University, Salisburylaan 133, 9820 Merelbeke, Belgium.

email: Luc.Duchateau@Ugent.be

and

Paul Janssen

Limburgs Universitair Centrum, Center for Statistics, Universitaire Campus, 3590 Diepenbeek, Belgium.

SUMMARY. Variance components are useful parameters to quantify the different sources of randomness in hierarchical models. In generalized mixed models and in frailty models a direct interpretation of the variance components is however not straightforward. A better approach is to see how the estimated values of the variance components affect the variability of specific quantities of interest. Depending on the concrete study such quantities might be the prevalence or the median time to event. In this note we discuss different examples from veterinary science with clustering between animals and show for these examples how variance components can be interpreted in the case of a normal, binary and time to event response variable.

KEY WORDS: Logistic regression; Survival data; Random effects; Variance components.

1 Introduction

In recent years the use of hierarchical models has become the standard approach to describe the different sources of randomness typically present in complex experimental data sets, e.g. in randomised block designs it might be better to think about the block effects as a random effect rather than taking the classical approach in which the block effects are modelled as fixed effects parameters. Apart from the fact that mixed effects models describe the structure of the design in a more accurate way and that they use the available information in the data more efficient in the case of an unbalanced design (Duchateau and Janssen, 1999), the inferential procedures for mixed models also provide us with estimates of the variance components present in the mixed model under consideration. This is useful since these variance components are often more than just nuisance parameters (Carroll, 2003). For instance, in animal breeding the heritability coefficient is defined in terms of the variance components and therefore the variance components are actually the parameters of interest. Another example (Section 2) is the development of reference tables for particular measurements on animals. Also here it is the variability between animals together with the mean value of the measurement under consideration that allows the investigator to determine a standard range of values for normal animals. Even if the variance components are nuisance parameters, it is important to evaluate the remaining variability in the data after adjustment by covariates. In large epidemiological studies with many covariates, it is likely that one or more of these

covariates influence the outcome significantly due to high power to detect even small and maybe clinically meaningless effects. In such cases, it is important to assess the variability remaining in the data after covariate adjustment as the covariates might only explain a small part of the variability in the data (Section 3). In mixed models the interpretation of the estimates of the variance components is easy as the random effect operates on the same scale as the actual values. This is however not true in generalized mixed models and frailty models. For these models there is a need to interpret (the size) of the estimates of the variance components through their effects on quantities that have a biological and understandable meaning for the study at hand. In Section 3 we demonstrate this idea for a logistic mixed regression model with success proportion as quantity of interest. In Section 4 we consider frailty models and demonstrate how the variance component influences the median time to event (Duchateau et al., 2002) and the success proportion at a particular time point.

2 Heterogeneity in linear mixed models

Echographic measurements of the volume of the thyroid gland of dogs (beagles) can possibly be used to diagnose decreased thyroid gland functioning as diseases such as hypothyroidy (too low production of thyroid gland hormones) are associated with shrinking of the thyroid gland. Whether the echographic measurement is useful for this purpose largely depends on the variation in the volume of the thyroid gland

in the healthy dogs population as this variation will determine the range of values of the echographic measurements that can be considered to be normal. The aim of the study was therefore to establish reference tables to differentiate hypothyroid dogs (with shrunk thyroid glands) from healthy dogs. There are two measurements (left and right side) for each dog. Studying heterogeneity in data with normally distributed random error term and random effect is straightforward as the random error term and the random effect operate at the same scale and influence the observed value in a linear manner. Therefore, the variability in the data due to different random sources can be compared in the linear mixed model. The model is given by

$$y_{ij} = \mu + d_i + e_{ij} \tag{1}$$

with y_{ij} the volume for the j^{th} side of dog i ($i = 1, \ldots, 23$), μ the overall mean, d_i the random i^{th} dog effect and e_{ij} the random error term. It is assumed that all d_i's and e_{ij}'s are mutually independent and furthermore that $d_i \sim \text{N}(0, \sigma_d^2)$ and $e_{ij} \sim \text{N}(0, \sigma^2)$. The estimate of the overall mean is $\hat{\mu} = 0.415$, the estimates of the variance components are $\hat{\sigma_d^2} = 0.0103$ and $\hat{\sigma^2} = 0.0084$.

Thus, the variance of an individual measurement on a healthy dog (either left or right) is given by $\text{Var}\,(y_{i1}) = \text{Var}\,(y_{i2}) = \sigma_d^2 + \sigma^2$, and thus 90% of the measurements on healthy dogs will fall in the interval $\left[\mu - 1.645\sqrt{\sigma_d^2 + \sigma^2}; \mu + 1.645\sqrt{\sigma_d^2 + \sigma^2}\right]$. Substituting the variance component estimates leads to the interval [0.19;0.64].

When both sides would be assessed in the determination of the volume, the variance of the average would be given by $\text{Var}\left\{0.5(y_{i1} + y_{i2})\right\} = \sigma_d^2 + 0.5\sigma^2$ and the 90%

interval, after substituting the variance component estimates, by [0.22;0.61]. Considering the two measurements thus leads to a substantial reduction of the range of the 90% interval.

Investigators are sometimes interested as well in the proportion of the total variance explained by the dog factor which in this case is estimated to be

$$100 \times \frac{\hat{\sigma}_d^2}{\hat{\sigma}_d^2 + \hat{\sigma}^2} = 100 \times \frac{0.0103}{0.0103 + 0.0084} = 55\%$$

The proportion of the variability explained by the dog factor can be easily derived in the mixed model, which will no longer be true for the more complex models considered in Sections 3 and 4.

3 Heterogeneity in generalized mixed models

The interpretation of heterogeneity in generalized mixed models based on the estimated variance components is more difficult because the random effect influences the observations in a non-linear way. Furthermore, the response measured on an observational unit, in our case study the number of animals infected, is binomially distributed whereas the random effect representing the cluster is assumed to come from a normal density. Therefore, it is difficult to compare the variance components of these two strata.

Interpretation of heterogeneity within the context of a generalized mixed model is studied in the following case study with binary outcome. Within each of 60 randomly selected farms, 30 pigs are individually screened for the presence of the

bacterial species Salmonella. This observational study was set up to investigate which factors at the farm level explain differences in prevalence over the different farms. The variance component attached to the farms can be considered to be a nuissance parameter, but we will demonstrate that it is worthwhile to study the variance component in the model with the covariates included as it gives us an idea of the remaining variability in the data.

First a model was fitted without covariates and with farm as random effect

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + f_i \tag{2}$$

with p_i the conditional proportion of infected pigs given the random farm effect f_i, μ the overall mean and f_i is the random effect of farm i. The distributional assumption is $f_i \sim \mathrm{N}\left(0, \sigma_f^2\right)$ and the f_i's are independent. We observe the number of pigs with Salmonella infection x_i with n_i pigs tested in farm i. The conditional distribution of x_i is binomial with parameters n_i and p_i, i.e. $x_i \sim \mathrm{B}\left(n_i; p_i \mid f_i\right)$.

Fitting this model gives $\hat{\mu} = 0.646$ and $\hat{\sigma}_f^2 = 7.74$. Thus, for a farm with random effect equal to 0, the conditional prevalence is estimated as 65.6% ($= \exp(\hat{\mu})/\{1 + \exp(\hat{\mu})\}$).

Adding covariates to the model we see that only the factor 'type of floor' explains the Salmonella prevalence in a significant way (p-value is 0.0052). We therefore propose the following extended model

$$\log\left(\frac{p_i}{1-p_i}\right) = \mu + \beta_1 x_{i1} + \beta_2 x_{i1} + f_i \tag{3}$$

with $x_{i1} = 1$ for fully slatted floor and 0 otherwise; $x_{i2} = 1$ for more than 50% but

not fully slatted floor and 0 otherwise; finally $x_{i1} = x_{i2} = 0$ corresponds to equal to or less than 50% slatted floor. Fitting this model gives $\hat{\mu} = 6.21$, $\hat{\beta}_1 = -6.06$, $\hat{\beta}_2 = -3.89$ and the variance estimate decreases to $\hat{\sigma}_f^2 = 4.52$.

The conditional prevalence for a farm with random effect equal to 0 and fully slatted floor is estimated as 53.8% (=$\exp(\hat{\mu} + \hat{\beta}_1)/\{1 + \exp(\hat{\mu} + \hat{\beta}_1)\}$), for more than 50% but not fully slatted floor as 91.1% and for less than 50% slatted floor as 99.8%. At this stage, it is important to interpret the variance components of the two models in order to assess whether the inclusion of the factor 'type of floor' leads to a substantial decrease of the variability of the prevalence over the different farms.

Since (2) implies $p_i = g(f_i) = \exp(\mu + f_i) / \{1 + \exp(\mu + f_i)\}$ we easily obtain (see Appendix) that

$$f_{p_i}(p) = \frac{1}{\sqrt{2\pi}\sigma_f} \exp\left[-\frac{1}{2\sigma_f^2}\left\{\log\left(\frac{p}{1-p}\right) - \mu\right\}^2\right] \frac{1}{p(1-p)} \tag{4}$$

Therefore the between farm variability σ_f^2 determines the shape of the prevalence density. The variance of the random effects that operate on the logit scale can be interpreted in terms of the spread of the prevalences over the farms by using the density function (4).

In order to compare the variability of the observed prevalences with the variability predicted from the logistic regression mixed model without covariates, the observed frequencies for 10 classes (0-10%, ..., 90-100 %) were derived and expected frequencies were obtained for these classes using the density function (4) with the estimates for μ and σ_f^2 inserted. The observed and expected probabilities for these 10 classes

are similar (Figure 1a).

The introduction of the covariate 'type of floor' decreases the variance at the level of the random effect in a substantial way from 7.74 to 4.52. The density function of the Salmonella prevalence for each floor type is also given by (4) with μ replaced by μ, $\mu+\beta_1$ or $\mu+\beta_2$ depending on the floor type. For the two categories without fully slatted floors, this leads to density functions that are concentrated around a high prevalence value with small probability to find farms in these two categories with low prevalence. The reduction of the variance, however, has little or no impact on the density function of the prevalence for the farms with fully slatted floor: prevalences in this type of farms still vary widely between 0 and 1 (Figure 1b).

In order to study the effect of different values of the variance component at different values for μ, the density function of p for a set of combinations of μ and σ_f^2 is shown in Figure 1c-d. Prevalences are more concentrated around a particular point when the variance of the random effect falls between 0.1 and 0.5 in the case that μ equals 0.646 (Figure 1c). With higher values for μ such as 2.2, this is true even for much higher values for the variance (Figure 1d). Thus the variance can not be assessed by itself but needs to be evaluated in conjunction with the overall prevalence.

4 Heterogeneity in frailty models

Also for survival data hierarchical models are needed for a good description of the design. As a case study, we investigate the time to first tick contact in cows. Cows

are clustered within farms leading to random farm effects. We first consider the frailty model without covariates

$$h_i(t) = u_i h_0(t) \tag{5}$$

with $h_i(t)$ the hazard of first contact at time t conditional on the frailty (random farm effect) u_i, $h_0(t)$ the baseline hazard of first contact at time t. We assume a one-parameter gamma density for u_i with mean 1 and variance σ_u^2

$$u_i \sim f_{u_i}(u) = \frac{u^{\frac{1}{\sigma_u^2}-1} \exp\left(-u/\sigma_u^2\right)}{\sigma_u^{2\,(1/\sigma_u^2)} \Gamma\left(1/\sigma_u^2\right)}$$

For cow j in farm i we observe the minimum of the censoring time c_{ij} and the time to first tick contact t_{ij}. We assume here an exponential model for the tick contact times, which is reasonable in the case that the tick challenge is constant in the period considered. The conditional distribution of the observed event times is then given by $t_{ij} \sim \exp\left(h_i \mid u_i\right)$ with $h_i = u_i h_0$.

We simulated data for 250 farms each with 90 cows that were followed up for 160 days with $h_0 = 0.01$ and $\sigma_u^2 = 0.04$. As a treatment factor, grazing system was considered: half the farms were supposed to use a communal grazing system, the other half a paddock grazing system with no or little contact between animals from different farms. A value of $\beta = -0.225$ was used in the simulation for the effect of the paddock grazing system.

Fitting the frailty model by the penalised likelihood approach (Klein and Moeschberger, 1997) gives $\hat{h}_0 = 0.00897$ and $\hat{\sigma}_u^2 = 0.053$. For a random farm effect equal to 1, the median time to contact is thus estimated as $\log(2)/\hat{h}_0$=77 days.

This model can be further extended with the grazing system covariate

$$h_i(t) = u_i h_0(t) \exp(\beta x_i) \tag{6}$$

where x_i is 0 for animals on communal grazing system farms and 1 for animals on paddock grazing system farms.

Fitting the extended model gives $\hat{h}_0 = 0.00979$, $\hat{\sigma}_u^2 = 0.044$ and $\hat{\beta} = -0.1933$. Thus, for a random farm effect equal to 1 the median time to first contact corresponds to 71 days for animals on communal grazing and to 86 days for animals on paddock grazing.

A direct interpretation of the estimated variance component, describing the between farm variability, is not easy since it operates on the hazard of first contact in a multiplicative way while the quantity of interest is the median time to first contact. Since, with m_i the median time to first contact for farm i, (5) implies $m_i = g(f_i) = \log(2)/(u_i h_0)$, we easily obtain (see Appendix) that

$$f_{m_i}(m) = \left(\frac{\log 2}{\sigma_u^2 h_0}\right)^{1/\sigma_u^2} \frac{1}{\Gamma(1/\sigma_u^2)} \left(\frac{1}{m}\right)^{1+\frac{1}{\sigma_u^2}} \exp\left(-\frac{\log 2}{\sigma_u^2 m h_0}\right) \tag{7}$$

In order to compare the variability of the Kaplan-Meier estimated median time to first contact from the different farms with the variability predicted from the frailty model without covariates, the observed frequencies for 16 classes were derived and expected frequencies were obtained for these classes using the density function (7) with the estimates for h_0 and σ_u^2 inserted. The observed frequencies correspond well with the expected frequencies from the frailty model (Figure 2a). The density functions for the median time to first contact for the two groups in the extended

model (6) do not only differ with respect to the mode but also with respect to the shape. Although both density functions are based on the same variance component estimate $\hat{\sigma_u^2}$, the density function is flatter for the group of herds with higher median time to contact, the farms with paddock grazing (Figure 2b). Thus, the effect of the variance of the random effect on the spread of the median event time also depends on the overal hazard rate, h_0 for the farms with communal grazing and $h_0 \exp(\beta)$ for the farms with paddock grazing.

Alternatively, the percentage of animals with first tick contact before a particular time point t, p_t, defined from the survival model as $p_t = 1 - \exp(-th_0u)$, can be used as a summary statistic to interpret the variance σ_u^2. The density function of the percentage of animals with first contact over the different farms at time t is given by (see Appendix)

$$f_{p_t}(p_t) = \frac{1}{th_0\sigma_u^{(1/\sigma_u^2)}\Gamma(1/\sigma_u^2)} (1-p_t)^{\frac{1}{th_0\sigma_u^2}-1} \left\{ \frac{-\log(1-p_t)}{th_0} \right\}^{\frac{1}{\sigma_u^2}-1} \tag{8}$$

Obviously, the density function of the percentage of animals with tick contact over the different farms is the flattest for that time point where the mode of the density function is close to 50%, corresponding to the percentage considered at 67 days (Figure 2c,d). The density function of the percentage of first tick contact at either an earlier (e.g. 30 days) or later time point (e.g. 150 days) is more concentrated as at those timepoints most farms have low, resp. high proportions of first tick contact.

5 Conclusions

Even if the variance seems to be a nuisance parameter, it is important in models with random effects to be able to interprete the value that the variance takes. Reporting the variance of the random effects in generalised mixed models and frailty models as such is not helpful as the actual value needs to be interpreted in terms of parameters that are easily understood by the investigator. Furthermore, even small absolute values for the random effects variance might lead to large variance in terms of the parameter of interest. For instance, Yamaguchi and Ohashi (1999) reported an estimated variance of a random effect in a frailty model of 0.01 as negligible, whereas it results in a 90% interval of median event times from 2 to 5 years. The same value for the variance might correspond to large variation in the parameter of interest in one model but not in another model, as the effect of the magnitude of the variance depends on the parameterisation and the mean value of the parameter of interest. Few diagnostic tools are available for survival models; even fewer exist for the random effects within the context of survival models. For instance, the best linear unbiased predictions of the random effects should be used with caution as they are shrinkage estimators and thus shrunk towards the mean and their variance will typically underestimate the variance (Morris, 1993). The reparameterisation given in this paper for the logistic regression mixed model and the frailty model allows us to depict the density function of the parameter of interest and compare it with the observed frequencies to assess whether the model assumptions seem reasonable. The

heterogeneity in the frailty model was interpreted in this paper assuming constant hazard rate. In most practical situations, however, the Cox proportional hazards model is used as the assumption of constant hazard rate can not be made and the models are then fitted by penalised partial likelihood (McGilchrist, 1993; Therneau et al., 2003). Currently we further investigate how to interprete the heterogeneity within the context of the Cox model.

Acknowledgements

The research of Luc Duchateau has been supported by the Fund for Scientific Research (grant no. 31520804) and the research of Paul Janssen by the Interuniversity Attraction Poles research network P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs).

References

Carroll, R.J. (2003). Variances are not always nuisance parameters. *Biometrics* **59**, 211-220.

Duchateau, L. and Janssen, P. (1999). An example-based tour in linear mixed models. In *Linear Mixed Models in Practice,* G. Verbeke, and G. Molenberghs (eds.), 11-61. Springer Verlag: New York.

Duchateau, L., Janssen, P., Lindsey, P., Legrand, C., Nguti, R., and Sylvester,

R. (2002). The shared frailty model and the power for heterogeneity tests in multicenter trials. *Computational Statistics and Data Analysis* **40**, 603-620.

Klein, J.P. and Moeschberger, M.L. (1997). *Survival analysis. Techniques for censored and truncated data.* Springer Verlag, New York.

McGilchrist, C.A. (1993). REML estimation for survival models with frailty. *Biometrics* **49**, 221-225.

Morris, J.S. (2002). The BLUPs are not "best" when it comes to bootstrapping. *Statistics and Probability Letters* **56**, 425-430.

Therneau, T.M., Grambsch, P.M. and Pankratz, V.S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics* **12**, 156-175.

Yamaguchi, T. and Ohashi, Y. (1999). Investigating centre effects in a multi-centre clinical trial of superficial bladder cancer. *Statistics in Medicine* **18**, 1961-1971.

APPENDIX

Derivation of (4). Consider the logistic regression model

$$\log\left(\frac{p}{1-p}\right) = \mu + f$$

where the random effect $f \sim f_f(f) = \mathrm{N}\left(0, \sigma_f^2\right)$. Since

$$p = \frac{\exp(\mu + f)}{1 + \exp(\mu + f)} = g(f)$$

is a montone transformation of f, we have

$$f_p(p) = f_f\left(g^{-1}(p)\right)\left|\frac{d}{dp}g^{-1}(p)\right|$$

and thus

$$f_p(p) = \frac{1}{\sqrt{2\pi}\sigma_f}\exp\left[-\frac{1}{2\sigma_f^2}\left\{\log\left(\frac{p}{1-p}\right) - \mu\right\}^2\right]\frac{1}{p(1-p)}$$

Derivation of (7) and (8). Consider the frailty model

$$h_(t) = uh_0(t)$$

with the assumption that $h_0(t) = h_0$, constant over time and u from a gamma density with mean 1 and variance σ_u^2

$$u \sim f_u(u) = \frac{u^{\frac{1}{\sigma_u^2}-1}\exp\left(-u/\sigma_u^2\right)}{\sigma_u^{2(1/\sigma_u^2)}\Gamma\left(1/\sigma_u^2\right)}$$

Here, interest is in the median time to event, i.e. $m = \log 2/h_0 u = s(u)$ which is again a montone transformation of u and thus we have

$$f_m(m) = f_u\left(s^{-1}(m)\right)\left|\frac{d}{dm}s^{-1}(m)\right|$$

leading to

$$f_m(m) = \left(\frac{\log 2}{\sigma_u^2 h_0}\right)^{1/\sigma_u^2} \frac{1}{\Gamma(1/\sigma_u^2)} \left(\frac{1}{m}\right)^{1+\frac{1}{\sigma_u^2}} \exp\left(-\frac{\log 2}{\sigma_u^2 m h_0}\right)$$

For this frailty model, the relationship between the percentage failure before a particular time point t, p_t, and u is $p_t = 1 - \exp(-th_0u) = z(u)$ which is again a monotone transformation with

$$f_{p_t}(p_t) = f_u\left(z^{-1}(p_t)\right)\left|\frac{d}{dp_t} z^{-1}(p_t)\right|$$

leading to

$$f_{p_t}(p_t) = \frac{1}{th_0\sigma_u^{(2/\sigma_u^2)}\Gamma(1/\sigma_u^2)} (1-p_t)^{\frac{1}{th_0\sigma_u^2}-1} \left\{\frac{-\log(1-p_t)}{th_0}\right\}^{\frac{1}{\sigma_u^2}-1}$$

Figure 1: The density function for the prevalence over different farms in the logistic mixed regression model. Observed (bars) and expected (circles) frequencies are shown in (a), the density functions of the prevalence for the three different types of floor are shown in (b). Finally, the density functions for different values of the variance are shown with μ equal to 0.645 (c) and 2.19 (d).

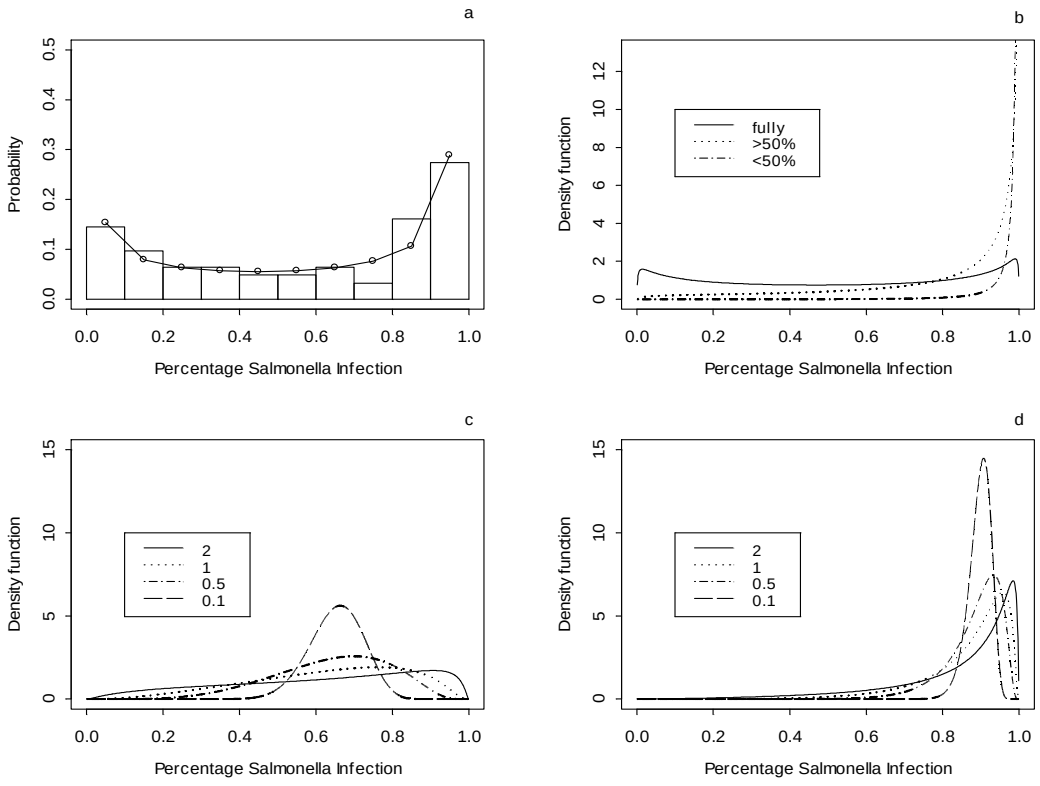


Figure 2: The density function for the median time to insemination and success proportion at a particular time point over different farms in the frailty model. Observed (bars) and expected (circles) frequencies of median event times are shown in (a), the density functions of the median for the two types of farms in (b). Finally, the density functions of the proportion of animals with tick contact at different time points (see legend: time points 30, 67, 150 and 190 days considered) are shown for the farms with free roaming animals ($h_0 = 0.00979$) (c) and for farms with herded animals ($h_0 \exp(\beta) = 0.00807$)(d).