NOTE TO USERS

University of Alberta

BOOTSTRAP METHODS FOR ONE-WAY RANDOM EFFECTS MODEL: APPLICATIONS TO INFERENCE ON NUMBER OF EXCEEDANCES

by

Qiaohao Zhu

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of
Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
Edmonton, Alberta
Fall, 2005

Your file Votre référence
ISBN: 0-494-09333-1
Our file Notre référence
ISBN: 0-494-09333-1

DEDICATION

To

My wife, Wen, who is always supporting on my study,

My daughter, Grace, who gives me so much pleasure ever since she's born

ABSTRACT

In this thesis the one-way random effects model is used as an analysis tool for estimating the distributional behavior of large data values. We define number of exceedances as the number of observations that have data value larger than a given threshold. The distributional behavior of large data values is studied using three quantities: (i)expected number of exceedances, (ii)variance of number of exceedances, and (iii)probability of observing no exceedances. We first use a parametric method, following the framework of Solomon (1989). Under the assumptions that the random effects and the random errors are normally distributed, the above three quantities can be expressed as functions of the variance components. We then use bootstrap method to obtain robust estimates. Since classical bootstrap method is not appropriate under one-way random effects model, we propose three different bootstrap methods that are consistent for estimating the above mentioned three quantities. Our simulation study shows that the parametric method works best only when the normality assumption is met, and the third bootstrap method is very robust against the distributional assumption. An application to a real data set is also discussed.

ACKNOWLEDGEMENT

I wish to thank my supervisor, Dr. N.G.N. Prasad, for your guidance and inspiration throughout this journey, and for the many helpful and thoughtful comments during the preparation of this thesis. Your knowledge and inspiration have guided me to the diverse and interesting research areas in Statistics.

I wish to thank Dr. Shubhash Lele and Dr. Peter Blenis for taking time out of your schedules to be on my examination committee.

Thanks also go to Dr. Peter Hooper, Dr. Paul Wiens, Dr. K.C. Carrière, Dr. Rohana Karunamuni, and Dr. Ivan Mizera. It is always a very good experience to be sitting in your class.

Last, but most importantly, I wish to thank the Department of Mathematical and Statistical Sciences, University of Alberta, for providing me the opportunity to study here and the financial support.

Table of Contents

List of Figures

List of Tables

List of Notations

y_{ij}: the j^{th} observation in the i^{th} group,

a: number of groups,

n_i: number of observations in i^{th} group,

n: number of observations in each group under a balanced case,

N: $N = \sum_{i=1}^{a} n_i$,

$\bar{y}_{i.}$: the i^{th} group mean, $\bar{y}_{..}$: the overall mean of the observations in all the groups, $\bar{y}_{i.} = \frac{\sum_{j=1}^{n} y_{ij}}{n}$,

$\bar{y}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{N}$,

v_i: random effect associated with group i,

e_{ij}: random error associated with j^{th} observation in the i^{th} group,

μ: overall mean of the population where y_{ij} come from,

$\bar{v}_{.}$: $\bar{v}_{.} = \frac{\sum_{i=1}^{a} v_i}{a}$,

$\bar{e}_{i.}$: $\bar{e}_{i.} = \frac{\sum_{j=1}^{n} e_{ij}}{n}$,

$\bar{e}_{..}$: $\bar{e}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} e_{ij}}{N}$,

σ_v^2: variance of v_i,

σ_e^2: variance of e_{ij}.

$\hat{\mu}$: $\hat{\mu} = \bar{y}_{..}$,

$\hat{\sigma}_e^2$: ANOVA estimator of σ_e^2, $\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{a(n-1)}$,

$\hat{\sigma}_v^2$: ANOVA estimator of σ_v^2, $\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1} - \frac{\hat{\sigma}_e^2}{n}$,

Q_B: sum of squares between groups, $Q_B = \sum_{i=1}^{a} n_i\,(\bar{y}_{i.} - \bar{y}_{..})^2$,

Q_W: sum of squares within groups, $Q_W = \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2$,

$Q_{B(-i)}$: sum of squares between groups excluding the i^{th} group, $Q_{B(-i)} = \sum_{k \neq i} n_k (\bar{y}_{k.} - \bar{y}_{..})^2$,

$Q_{W(-i)}$: sum of squares within groups excluding i^{th} group, $Q_{W(-i)} = \sum_{k \neq i} \sum_{j=1}^{n} (y_{kj} - \bar{y}_{k.})^2$,

$\Delta = \frac{\sigma_v^2}{\sigma_e^2}$,

$\hat{\Delta} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2}$,

$\alpha = 1 - \sqrt{\frac{1}{1+n\Delta}}$,

$\hat{\alpha} = 1 - \sqrt{\frac{1}{1+n\hat{\Delta}}}$,

z_{ij}: data transformed from y_{ij}: $z_{ij} = y_{ij} - \alpha \bar{y}_{i.}$,

$\bar{z}_{..}$: overall mean of z_{ij}: $\bar{z}_{..} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{n} z_{ij}}{an}$,

z_k^*,z_{ij}^*: bootstrap sample drawn from z_{ij},

y_{ij}^*: data transformed from the bootstrap sample z_{ij}^*,

μ_z: population mean of z_{ij},

$\hat{\mu}_z^*$: bootstrap estimator of μ_z,

$\hat{\sigma}_e^{*2}$: bootstrap estimator of σ_e^2,

$\hat{\sigma}_v^{*2}$: bootstrap estimator of σ_v^2,

T_i: number of exceedances in the i^{th} group,

T: number of exceedances in each group, under a balanced case,

$E(T)$: expected number of exceedances,

$Var(T)$: variance of number of exceedances,

$Pr(T = 0)$: the probability of observing no exceedance,

$O(n^{-1})$ represents terms that have the same or smaller order of magnitude as n^{-1}, in the sense that $\left|\frac{O(n^{-1})}{n^{-1}}\right|$ is bounded as $a \to \infty$.

$O\left((a \wedge n)^{-1}\right)$ denotes $O(a^{-1})$ or $O(n^{-1})$ or both of them.

E_M, Var_M and Cov_M denote the expectation, variance an covariance with respect to the model.

E_*, Var_*, and Cov_* denote the expectation, variance and covariance with respect to the bootstrap sample.

Chapter 1

Introduction and Literature Review

Understanding the distributional behaviour of large data values is very useful in medical research, for example in studying number of times blood pressure values increase over a threshold value. We study the distributional behaviour of large data values under a one-way random effects model. Both parametric method and bootstrap methods will be considered in this thesis. In this chapter, we discuss an application of one-way random effects model to the estimate of number of exceedances and related quantities.

1.1 Introduction

Random effects models, also known as variance components models, have been widely used in many different fields of research. In epidemiologic research, they are commonly used to measure the degree of familial resemblance with respect to biological characteristics. In genetics these models play a central role in estimating the heritability of selected traits in animal and plant populations.

In sample surveys, statisticians use random effects models to improve the estimate accuracy for the areas or population groups that do not have enough representative sample sizes. The simplest case of a random effects model is a one-way random effects model (see Searle et al, 1992). Under the one-way random effects model all data observations are classified into groups according to some criteria, for example, all the measurements from each subject are classified into one group. The groups in the data are assumed to be a random sample of groups from all possible groups. For example, if groups are subjects, then, all the subjects included in the data are assumed to be a finite sample from the subject population. There are two different sources of variations under one-way random effects model: one is the variation of observing only a sample of groups, which is called the group effect, or random effect, associated with each group of observations, the other is the variation of measurement, called random error, which is associated with each observation. The primary interest in the random effects model is to estimate the two sources of variations.

We use the data set from Solomon (1989) in this thesis. Solomon (1989) considers a data set on blood pressure from the International Prospective Primary Prevention Study in Hypertension (IPPPSH). The data are the observations made quarterly for a period of 4 years (thus a total of 16 measurements from each patient) on 25 hypertensive males receiving treatment regimens containing a betablocker, the measured variables are diastolic and systolic pressures. Using this data set, we want to find out how many blood pressure measurements from a patient will be higher than a given threshold, that is, we are interested in estimating the behaviour of large blood pressure measurements.

The data on blood pressure can be classified in the following way: the measurements from a patient form a group, giving a total of 25 groups, and within each group, we have 16 observations. Thus the total variation of the blood pressure measurements is consisted of two components: between groups (patients) and within group (each patient). We then use one-way random effects model to analyze this data set.

The one-way random effects model can be described as:

$$y_{ij} = \mu + v_i + e_{ij}, \qquad j = 1, \ldots, n_i, \qquad i = 1, \ldots, a, \tag{1.1}$$

where y_{ij} is the value associated with the j^{th} observation for the i^{th} group (such as the j^{th} blood pressure measurement from the i^{th} patient), μ is the overall mean, v_i is the random effect associated with i^{th} group, and e_{ij} is the random error associated with the j^{th} observation in the i^{th} group. Here v_i and e_{ij} are assumed identically independent random variables, with mean 0 and variances σ_v^2 and σ_e^2, respectively, and possibly normally distributed. Further, v_i and e_{ij} are assumed to be independently distributed. Here, σ_v^2 and σ_e^2 are unknown parameters, which are called variance components.

If all $n_i = n$ for $i = 1, 2, \ldots, a$, then the model is called balanced one-way random effects model.

In the IPPPSH blood pressure data, y_{ij}'s represent the blood pressure measurements, v_i describes the effect of the i^{th} patient, and e_{ij} describes the measurement error of taking j^{th} blood pressure for the i^{th} patient. v_i and e_{ij} are not observed.

Note that the above one-way random effects model has three unknown parameters, μ, σ_v^2, and σ_e^2. In the classical framework of random effects models (or mixed effects models), the primary interests are on the estimation of μ and the variance components or functions of variance components. For this purpose, many methods are available for general random and mixed effects models. These methods include: Henderson's Method I, Method II, and Method III (Henderson (1953)), Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML), Bayes estimations, and Minimum Norm Quadratic Estimation (MINQUE, Rao, 1970, 1971, 1972). ANOVA type of methods, which includes Henderson's three methods, do not require the assumption of normality. Searle et al (1992) gives detailed treatment on the estimation of variance components.

For the simple one-way random effects model, the estimation of the two variance components is very simple, especially under a balanced case. We

discuss this in Chapter 2. In this thesis, our primary purpose is not on the estimation of the variance components for the model, but on the estimation of the distributional behaviour of number of large data values, which is not directly described by the one-way random effects model. Solomon (1989) first discussed this problem. She modeled the IPPPSH data set using a one-way random effects model. Under the normality assumptions of the random effects and the random errors, the distribution of the number of large data values can be easily derived from model (1.1).

Under model (1.1), if we define the large data value as an indicator function for the observation y_{ij} as:

$$I_{ij}(h) = \begin{cases} 1 & y_{ij} > h \\ 0 & \text{otherwise} \end{cases}$$

where h is a given threshold. Then, we can define the number of large data values, which we termed as number of exceedances T_i, for the i^{th} group, as the summation of $I_{ij}(h)$ over all observations in the i^{th} group. That is,

$$T_i = \sum_{j=1}^{n_i} I_{ij}(h)$$

Since T_i is a function of y_{ij}, for $j = 1, 2, \ldots, n_i$, which are normally distributed (under the normality assumptions of random effects and random errors), we can then easily derive the distribution of T_i. In a balanced case, all T_i have identical distributions, in which case we drop the subscript i. Solomon (1989) considers three parameters based on T as the interested quantities. These three quantities are: (a) $E(T)$, the expected number of exceedances, (b) $Var(T)$, the variance of number of exceedances, and (c) $Pr(T = 0)$, the probability of observing no exceedance. These three quantities can be expressed as functions of the model parameters: the overall mean μ, the variance of random effects σ_v^2, and the variance of random errors σ_e^2, as follows:

$$E(T) = n\Phi\left(\frac{\mu - h}{\sigma}\right), \tag{1.2}$$

$$Var(T) = n\Phi\left(\frac{\mu - h}{\sigma}\right)\{1 - \Phi\left(\frac{\mu - h}{\sigma}\right)\}[1 + (n-1)\rho_I], \tag{1.3}$$

and,

$$Pr(T=0) = \int_{-\infty}^{+\infty} \phi(x) \left[\Phi\left(\frac{\gamma - x}{\tau}\right)\right]^n dx, \tag{1.4}$$

where

$$\rho_I = \frac{\Phi_2\left(\frac{\mu-h}{\sigma}, \frac{\mu-h}{\sigma}; \rho = \frac{\sigma_v^2}{\sigma^2}\right) - \Phi^2\left(\frac{\mu-h}{\sigma}\right)}{\Phi\left(\frac{\mu-h}{\sigma}\right)\left[1 - \Phi\left(\frac{\mu-h}{\sigma}\right)\right]},$$

$$\gamma = (h-\mu)/\sigma_v,$$

$$\tau = \sigma_e/\sigma_v,$$

and

$$\sigma^2 = \sigma_v^2 + \sigma_e^2,$$

with $\phi(x)$ and $\Phi(x)$ denoting the standard normal probability density function and cumulative density function, respectively.

The three interested quantities all depend on the model parameters μ, σ_v^2, and σ_e^2, which are unknown. We can use the ANOVA method (such as Henderson's three methods) or Maximum Likelihood method to estimate them from the observed data y_{ij}, and then plug these estimates in the above formulae (1.2), (1.3), and (1.4) to obtain the point estimates of $E(T)$, $Var(T)$ and $Pr(T=0)$.

As discussed in Chapter 2, the form of these functions are quite complicated, making it almost impossible to estimate the variances of these estimates, and further, these functions depend heavily on the assumption of normality, and thus they are not robust estimators. In order to address these problems, we propose bootstrap methods for the estimation of these three quantities.

The main difficulty in applying classical bootstrap methods to this problem is that these methods are developed for the identically independently distributed (i.i.d.) data. Under the i.i.d. data case, they are consistent and robust. But under one-way random effects model, the observed data y_{ij} are not i.i.d, so we can not directly apply the classical bootstrap methods. We have to consider new bootstrap methods that are suitable for this non-i.i.d case. These procedures are described in Chapter 3.

1.2 Literature Review for the Theory of Bootstrap

The main purpose of this thesis is to estimate the three quantities proposed in Solomon (1989), using bootstrap procedures, so in this section, we review some development of the bootstrap theory that are relative to our problem. For the one-way random effects model, there are great amount of literature. In particular, the book on variance components by Searle et al (1992) is very useful.

Since the introduction of bootstrap by Efron (Efron, 1979), with the rapid development of fast computing ability, the bootstrap method has become an intensively used method for assessing uncertainty in a vast range of domains, from i.i.d. case to independent case to correlated models.

The bootstrap method can be viewed as a mixture of two techniques: the substitution principle and the Monte Carlo method for numerical approximation (Shao & Tu, 1995).

Often parameters can be expressed (implicitly or explicitly) as a function of the underlying distribution:

$$\theta = R_n(X_1, X_2, \ldots, X_n) \tag{1.5}$$

$$(X_1, X_2, \ldots, X_n) \sim P_n \tag{1.6}$$

where R_n is a function of $X_1, X_2, \ldots, X_n$, and the data $X_1, X_2, \ldots, X_n$, are generated from the distribution P_n (possibly from a distribution family, but the exact distribution is unknown), thus the form of the function R_n depends on P_n.

The traditional approach of finding an estimate of θ, say, $\hat{\theta}$, is to use P_n to find a function R_n, and then plug in the observed data $X_1, X_2, \ldots, X_n$. But sometimes it is difficult or even impossible to find a closed form of R_n, and furthermore, it will be even more difficult to find the accuracy measures (such as variance, mean squared error) of $\hat{\theta}$.

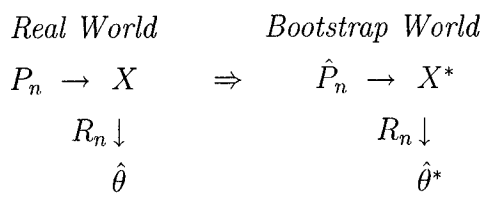

Figure 1.1: Typical Bootstrap Diagram

The bootstrap approach instead tries to estimate the distribution P_n, denoted as $\hat{P}_n$, from the observed data $X_1, X_2, \ldots, X_n$, and then plug in (1.6) (substitution principle). Since now $\hat{P}_n$ is completely known, we can generate as many data $X_1^*, X_2^*, \ldots, X_n^*$ as we want from $\hat{P}_n$, and for each generated data we compute an estimate from (1.5), say $\hat{\theta}^*$ (Monte Carlo method). Thus we have many $\hat{\theta}^*$, then we can have their mean value and the variance (or mean squared error), which are used as the bootstrap estimate and the accuracy measure of the interested parameter θ. The general bootstrap approach is described in figure 1.1 (Efron, 2003).

This figure shows that from Real World to Bootstrap World, we simply replace P_n with its point estimate $\hat{P}_n$. Usually we have two different approaches for estimating $\hat{P}_n$ from the given observed data $X = (X_1, X_2, \ldots, X_n)$. One is to assume P_n comes from a distribution family, depending on some unknown parameters. We first estimate these unknown parameters from the observed data $X_1, X_2, \ldots, X_n$, and then plug in P_n to get $\hat{P}_n$, and the bootstrap samples are generated from $\hat{P}_n$. This is called a parametric bootstrap. Another approach is that we do not assume any particular form of P_n, instead we use the empirical distribution $\hat{F}_n$ from $X_1, X_2, \ldots, X_n$, which assigns $\frac{1}{n}$ of probability to each X_i, for $\hat{P}_n$. This is called a nonparametric bootstrap. For the nonparametric bootstrap method, since we do not assume any specific form of the underlying distribution, it is thus distributionally robust, and is widely used.

From the above diagram we see that if the "Bootstrap World" really mimics the "Real World", then the bootstrap method will provide a better result.

In this sense, the bootstrap method is actually model-dependent: the "Bootstrap World" should fairly reflect the "Real World", or the real model, in other words. But the bootstrap method requires no theoretical formula for the quantity to be estimated and thus it is less model-dependent than the traditional approach.

The difference between the "Bootstrap World" and the "Real World" is determined by the two distribution P_n and $\hat{P}_n$, which are the data generation mechanisms. Since in practice, we often use the empirical distribution $\hat{F}_n$ for $\hat{P}_n$, thus it is intuitively correct that this bootstrap method works better when the observed data $X_1, X_2, \ldots, X_n$ are independently identically distributed (i.i.d.) from P_n. This is well addressed and justified in literature. The first two important papers on the asymptotic accuracy are by Singh (1981) and Bickel and Freedman (1981), which showed that the bootstrap procedure can deliver higher-order accuracy than the approximation by the limiting normal distribution for statistics that can be expressed as functions of sample means.

It is also well known that extending this classical bootstrap procedure to independent but not identically distributed data, or correlated data is quite difficult, and may not attain the consistency and asymptotic accuracy.

In the context of linear models, several algorithms have been developed for the bootstrap method (Shao & Tu, 1995). The first algorithms, external bootstrap (EB) or external weighted bootstrap, is proposed by Wu (1986) for the least squared estimator (LSE) in linear regression. For the regression model: $y_i = x_i'\beta + \epsilon_i$, let β_{LS} be the ordinary least squared estimator of β, and $\hat{\epsilon}_i$ be the residuals after fitting the model. Let t_i^* be i.i.d. from a distribution with mean 0 and variance 1, then the bootstrap samples are generated by setting:

$$y_i^* = x_i'\hat{\beta}_{LS} + \frac{|\hat{\epsilon}_i|}{\sqrt{1-h_i}} t_i^* \tag{1.7}$$

and $\hat{\beta}^*$, the bootstrap estimator of β, is the ordinary least squared estimator based on the data (y_i^*, x_i'). Here the bootstrap estimator depends on t_i^*, which is independent of the original data (y_i, x_i'). Liu (1988) provided some extensions

and gave theoretical justification for Wu's EB procedure, and suggested having another restriction on t_i^*: $E_*(t_i^{*3}) = 1$. Our first bootstrap method is based on this procedure.

The second procedure is bootstrapping residuals (RB), which can be viewed as semi-parametric bootstrap. This is done by first identifying the linear model by a parameter, such as β in the regression model $Y = X\beta + \varepsilon$, then find an estimate for the parameter $\hat{\beta}$, and then computing the residuals $\hat{\epsilon}_i$, where the bootstrap samples are drawn from. This seems to be intuitive since in general, we assume the error term in the model is i.i.d., if the model is correct and the estimate of the parameter is very close to the true value, then the residuals after fitting the model will have the same distribution as the error term, which is then i.i.d. (approximately), and bootstrap method can be applied to the almost identically independently distributed residuals. Cautions should be made here that, first, the fitted residuals $\hat{\epsilon}_i$ are not independent, they are actually correlated, second, the usefulness of this procedure depends on the model assumption, if the model is incorrect, then the fitted residuals $(\hat{\epsilon}_i)$ will not distribute approximately the same as the error terms (ε), and third, the estimated parameter should be consistent to the true parameter value. Our second and third bootstrap methods are the extension the RB procedure.

The third procedure is called paired bootstrap (PB), where the bootstrap samples are generated from (y_i, x_i'), $i = 1, 2, \ldots, n$. In this case, the model can be identified by the joint distribution of (y_i, x_i'), and estimated by the empirical distribution, putting mass $\frac{1}{n}$ to each (y_i, x_i').

A more recently developed bootstrap methodology for the non-i.i.d. models is bootstrapping estimating functions, the combination of bootstrap method and the theory of estimating function. The idea of this methodology is that the unknown parameters can be expressed as a solution to a series of estimating functions. We then treat each estimating function as if they were the observed data, and apply the classical bootstrap method on the estimating functions. The justification of this method is: the expected value of each estimating function is 0, so they have the same mean, although their variance may be

different, but by Liu (1988), under some mild conditions, the bootstrap method applied in this case retains the same robust properties as when it is applied under the i.i.d. case. The method of bootstrapping estimating functions is robust and requires less computation. Lele (2003) gives an excellent review of this methodology, and a good example of this method in regression context can be found in Hu and Zidek (1995).

Chapter 2

Parametric method

In this Chapter, we describe the parametric method developed by Solomon (1989) for estimating the distributional behaviour of the number of exceedances over a threshold under one-way random effects model with normality assumption, and present the analytical formulae for the three interested quantities: the expected number of exceedances, the variance of the number of exceedances and the probability of observing no exceedances.

2.1 One-way Random Effects Model

The one-way random effects model is described as below:

$$y_{ij} = \mu + v_i + e_{ij}, \qquad j = 1, \ldots, n_i, \qquad i = 1, \ldots, a, \tag{2.1}$$

with

$$v_i \overset{iid}{\sim} N(0, \sigma_v^2) \qquad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2).$$

and further, $\{v_i\}$ and $\{e_{ij}\}$ are independently distributed. Here, y_{ij} is j^{th} observation in the i^{th} group, μ is the overall mean and is a constant value. v_i is the random effect associated with group i, and e_{ij} is the random error

associated with the observation of y_{ij}, this usually is the measurement error. μ, σ_v^2, and σ_e^2 are unknown parameters.

Both the random effects $\{v_i\}$ and the random errors $\{e_{ij}\}$ are assumed to be independently, identically distributed with mean 0 and variances σ_v^2 and σ_e^2, respectively, and also, v_i and e_{ij} are independent, for all $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, n_i$. Thus the variance-covariance structure of y_{ij} is given by

$$cov(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_v^2 + \sigma_e^2 & \text{for } i = i' \text{ and } j = j' \\ \sigma_v^2 & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

The unknown parameter μ is estimated by the overall sample mean $\bar{y}_{..} = \frac{1}{\sum_{i=1}^{a} n_i} \sum_{i=1}^{a} \sum_{j=1}^{n_i} y_{ij}$, and the two variance components σ_v^2 and σ_e^2 are usually estimated by the classical ANOVA estimation method. We describe the method as below (Searle et al, 1992):

Let SSA be the sum of squares between groups, SSE be the sum of squares within groups, that is:

$$SSA = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2, \tag{2.3}$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2, \tag{2.4}$$

and let

$$MSA = \frac{SSA}{a-1}, \tag{2.5}$$

$$MSE = \frac{SSE}{a(n-1)}. \tag{2.6}$$

then, the ANOVA estimators of σ_v^2 and σ_e^2 are, respectively,

$$\hat{\sigma}_v^2 = \frac{MSA - MSE}{(N - \sum_i n_i^2/N)/(a-1)} \tag{2.7}$$

and

$$\hat{\sigma}_e^2 = MSE \tag{2.8}$$

where $N = \sum_{i=1}^{a} n_i$.

For the balanced data, *i.e.*,when $n_1 = n_2 = \ldots = n_a$, numbers of observations for each group are the same, the ANOVA estimators of σ_v^2, σ_e^2 can be simplified to:

$$\hat{\sigma}_v^2 = \frac{1}{n}(MSA - MSE) \tag{2.9}$$

$$\hat{\sigma}_e^2 = MSE \tag{2.10}$$

One advantage of using the ANOVA estimator is that, it does not require the normal distributional assumptions for the random effects and the random errors, and they have the unbiasness and consistency properties. But there are cases where $\hat{\sigma}_v^2$ maybe negative. In this case, we may simply let $\hat{\sigma}_v^2 = 0$ and $\hat{\sigma}_e^2 = \frac{SSA+SSE}{N-1}$. We can also use the Maximum Likelihood Estimators for estimating σ_v^2 and σ_e^2, which will guarantee that we will get the positive estimates for the two variance components, but the maximum likelihood estimators are not unbiased. We will use the ANOVA estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ in this thesis.

2.2 Distributional Behaviour of Large Data Values

The main interest in this thesis is the estimation of the distributional behaviour of large data values under one-way random effects model. Following Solomon (1989), we use three quantities to describe the behaviour of large data values: The expected value of the number of exceedances, the variance of number of exceedances, and the probability of observing no exceedance. We give their definitions in this section.

First, we define **Number of Exceedances** for the i^{th} group: Let h be a given threshold value, define:

$$I_{ij}(h) = \begin{cases} 1 & y_{ij} > h \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

Then, the number of exceedances for the i^{th} group is the sum of n_i random variables:

$$T_i = \sum_{j=1}^{n_i} I_{ij}(h) \tag{2.12}$$

From model (2.1), we have $y_{ij} \sim N(\mu, \sigma_v^2 + \sigma_e^2)$, thus $I_{ij}(h)$ is Bernouli distributed, with success probability $p = Pr(y_{ij} > h)$. then, T_i is the sum of n_i Bernouli random variables, but T_i is not necessarily binomial, because $I_{ij}(h)$ are not independent.

Since T_i is a random variable, the three interested quantities are then defined as:

(1) **The Expected number of exceedances**: The expectation of T_i: $E(T_i)$;

(2) **The variance of the number of exceedances**: The variance of T_i: $Var(T_i)$;

(3) **The probability of observing no exceedance**: The probability of $T_i = 0$: $Pr(T_i = 0)$.

Here, T_i are group specific. But following Solomon (1989), we only use the balanced one-way random effects model for our study, that is, the number of observations within each group are the same. Hence, the distribution of T_i will be identical for all i, so we drop the subscribe i for simplicity in the following descriptions.

Given the observed data y_{ij}, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n$, according to the above definition, we can compute the observed $E(T)$, $Var(T)$ and $Pr(T = 0)$ as below:

$$E_{obs}(T) = \frac{\sum_{i=1}^{a} T_i}{a} \tag{2.13}$$

$$Var_{obs}(T) = \frac{1}{a-1} \sum_{i=1}^{a} (T_i - E_{obs}(T))^2 \tag{2.14}$$

$$Pr_{obs}(T = 0) = \frac{\sum_{i=1}^{a} \prod_{j=1}^{n} I_{ij}(h)}{a} \tag{2.15}$$

2.3 Parametric Method

In this section, we derive the parametric formulae for the estimation of the three quantities according to Solomon (1989), based on the balanced one-way random effects model, under normality assumptions: $v_i \sim N(0, \sigma_v^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n$.

Let $\sigma^2 = \sigma_v^2 + \sigma_e^2$, from model (2.1), we have:

$$y_{ij} \sim N(\mu, \sigma^2) \tag{2.16}$$

Then, the indicators $I_{ij}(h)$ are correlated binomial random variables where:

$$\begin{aligned} E[I_{ij}(h)] &= Pr(y_{ij} > h) \\ &= Pr(\frac{y_{ij} - \mu}{\sigma} > \frac{h - \mu}{\sigma}) \\ &= \Phi\left(\frac{\mu - h}{\sigma}\right); \end{aligned}$$

$$\begin{aligned} Var[I_{ij}(h)] &= E[I_{ij}(h)]\{1 - E[I_{ij}(h)]\} \\ &= \Phi\left(\frac{\mu - h}{\sigma}\right)\left[1 - \Phi\left(\frac{\mu - h}{\sigma}\right)\right]; \end{aligned}$$

and for $j \neq k$:

$$\begin{aligned} Cov[I_{ij}(h), I_{ik}(h)] &= E[I_{ij}(h)I_{ik}(h)] - E[I_{ij}(h)]E[I_{ik}(h)] \\ &= Pr(y_{ij} > h, y_{ik} > h) - \Phi^2\left(\frac{\mu - h}{\sigma}\right) \\ &= Pr\left(\frac{y_{ij} - \mu}{\sigma} > \frac{h - \mu}{\sigma}, \frac{y_{ik} - \mu}{\sigma} > \frac{h - \mu}{\sigma}\right) - \Phi^2\left(\frac{\mu - h}{\sigma}\right) \\ &= \Phi_2\left(\frac{\mu - h}{\sigma}, \frac{\mu - h}{\sigma}; \rho = \frac{\sigma_v^2}{\sigma^2}\right) - \Phi^2\left(\frac{\mu - h}{\sigma}\right) \end{aligned}$$

with $\Phi_2(x, y; \rho)$ representing the standardized bivariate normal distribution function with correlation ρ.

Following the above results,

$$E(T) = E\left[\sum_{j=1}^{n} I_{ij}(h)\right] = nE[I_{ij}(h)] = n\Phi\left(\frac{\mu - h}{\sigma}\right) \tag{2.17}$$

and

$$\begin{aligned} Var(T) &= Var\left[\sum_{j=1}^{n} I_{ij}(h)\right] \\ &= \sum_{j=1}^{n} Var\left[I_{ij}(h)\right] + \sum_{j\neq j'} Cov\left[I_{ij}(h), I_{ij'}(h)\right] \\ &= nVar\left[I_{ij}(h)\right] + n(n-1)Cov\left[I_{ij}(h), I_{ij'}(h)\right] \\ &= n\Phi\left(\frac{\mu-h}{\sigma}\right)\left[1-\Phi\left(\frac{\mu-h}{\sigma}\right)\right] \\ &+ n(n-1)\left[\Phi_2\left(\frac{\mu-h}{\sigma}, \frac{\mu-h}{\sigma}; \rho=\frac{\sigma_v^2}{\sigma^2}\right) - \Phi^2\left(\frac{\mu-h}{\sigma}\right)\right] \\ &= n\Phi\left(\frac{\mu-h}{\sigma}\right)\{1-\Phi\left(\frac{\mu-h}{\sigma}\right)\}[1+(n-1)\rho_I] \end{aligned} \tag{2.18}$$

where

$$\rho_I = \frac{\Phi_2\left(\frac{\mu-h}{\sigma}, \frac{\mu-h}{\sigma}; \rho=\frac{\sigma_v^2}{\sigma^2}\right) - \Phi^2\left(\frac{\mu-h}{\sigma}\right)}{\Phi\left(\frac{\mu-h}{\sigma}\right)\left[1-\Phi\left(\frac{\mu-h}{\sigma}\right)\right]}$$

To find the probability of no exceedances, $Pr(T=0)$, note that given v_i:

$$y_{ij}|v_i \sim N(\mu+v_i, \sigma_e^2)$$

and,

$$v_i \sim N(0, \sigma_v^2)$$

we have:

$$Pr(T=0) = \int_{-\infty}^{+\infty} Pr(T_i=0|v_i)f(v_i)dv_i$$

where

$$\begin{aligned} Pr(T=0|v_i) &= Pr(y_{i1}|v_i \leq h, \ldots, y_{in}|v_i \leq h) \\ &= [Pr(y_{ij}|v_i \leq h)]^n \\ &= \left[\Phi\left(\frac{h-\mu-v_i}{\sigma_e}\right)\right]^n \end{aligned}$$

and $f(v_i)$ is the density function of v_i, which is:

$$f(v_i) = \frac{1}{\sigma_v}\phi\left(\frac{v_i}{\sigma_v}\right).$$

Then

$$\begin{aligned} Pr(T=0) &= \int_{-\infty}^{+\infty} \frac{1}{\sigma_v}\phi\left(\frac{v_i}{\sigma_v}\right)\left[\Phi\left(\frac{h-\mu-v_i}{\sigma_e}\right)\right]^n dv_i \\ &= \int_{-\infty}^{+\infty} \phi(x)\left[\Phi\left(\frac{h-\mu-x\sigma_v}{\sigma_e}\right)\right]^n dx \\ &= \int_{-\infty}^{+\infty} \phi(x)\left[\Phi\left(\frac{\gamma-x}{\tau}\right)\right]^n dx \end{aligned} \tag{2.19}$$

where $\gamma = (h-\mu)/\sigma_v$, $\tau = \sigma_e/\sigma_v$, and $\phi(x)$ and $\Phi(x)$ are the standard normal probability density function and cumulative density function, respectively.

The equations (2.17),(2.18), and (2.19) are all functions of the unknown parameters μ, σ_v^2 and σ_e^2. We can use the method described in Section 2.1 to find the estimates of the three model parameters, and then plug these estimates in the equations to estimate $E(T)$, $Var(T)$ and $Pr(T=0)$.

To find the sampling variances of $E(T)$, $Var(T)$ and $Pr(T=0)$, we can use the delta method, since they are smooth functions of the three model parameters, μ, σ_v^2 and σ_e^2. But this is not easy, since it will be very difficult to find the derivative functions from (2.17), (2.18), and (2.19) with respect to the three parameters, thus making variance estimation quite difficult.

Chapter 3

Proposed Bootstrap Methods

We propose three bootstrap methods for the estimation of the distributional behaviour of the number of exceedances in this Chapter. The first bootstrap method is based on Wu's external weighted bootstrap procedure, which draws bootstrap weights from an external distribution that are not related to the observed data. The idea behind the second bootstrap method is to transform the non i.i.d. data to almost i.i.d. data, and then apply the classical bootstrap procedure on the transformed data to estimate the model parameters, and the three interested quantities can be estimated by plugging the model parameters in the formulae from Solomon's parametric method. The third bootstrap method differs from the second bootstrap method in the way that after drawing the bootstrap samples from the transformed data, we apply the inverse transformation on the bootstrap samples, and compute the estimates of the three interested quantities directly from them. For each of the three bootstrap method, we give a description of the procedure and show that the estimates are consistent for the interested quantities.

3.1 Introduction

Using the formulae of (2.17), (2.18), and (2.19), and the ANOVA or Maximum Likelihood estimates of μ, σ_v^2 and σ_e^2, we can get the parametric estimates of the expected number of exceedances and its variance, and the probability of observing no exceedance. There are several disadvantages for the parametric method.

The first disadvantage is that, the parametric method relies heavily on the assumption of normality. This can be seen from the formulae of (2.17), (2.18), and (2.19). $E(T)$, $V(T)$ and $Pr(T = 0)$ are all explicit functions involving the cumulative density function (cdf) of the standard normal distribution ($\Phi(x)$), thus the validity of these estimators relies on the normality assumption. In cases when the data are not coming from normal distribution, the estimates of $E(T)$, $V(T)$ and $Pr(T = 0)$ will not be correct.

The second disadvantage is that, it is very difficult to derive explicit expressions of the variance for these estimators. If we treat $E(T)$, $V(T)$ and $Pr(T = 0)$ as functions of μ, σ_v^2 and σ_e^2, and apply the delta-method, it is possible to compute their approximate variance estimates. But from (2.17), (2.18), and (2.19), we see that these estimators are complicated functions of μ, σ_v^2 and σ_e^2, especially for the estimator of $Pr(T = 0)$. It is not easy to find the partial derivatives with respect to μ, σ_v^2 and σ_e^2, thus, it is difficult to apply the delta-method.

In order to overcome the above disadvantages of the parametric methods, we use a bootstrap method. In general, unlike traditional parametric approaches, the bootstrap procedures do not require a theoretical form of the distribution underlying the data, thus they are robust against distributional assumptions. However, although the bootstrap method has drawn a great deal of attention in recent years, most of the theoretical work is for the independent and identically distributed (i.i.d.) cases; its applicability to cases other than

i.i.d. is not justifiable in general. In our case, for the model (1.1), we have

$$cov(y_{ij}, y_{i'j'}) = \begin{cases} \sigma_v^2 + \sigma_e^2 & \text{for } i = i' \text{ and } j = j' \\ \sigma_v^2 & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise} \end{cases} . \tag{3.1}$$

Thus the data $\{y_{ij}\}$ are not i.i.d, we can not directly apply the results of classical bootstrap procedure to the observed data $\{y_{ij}\}$. How to develop suitable bootstrap methods for one-way random effects model and compare their performances is the main objective of this thesis. In this Chapter, we present several bootstrap methods that are applicable for our model, and in the following chapter we compare their performances through a simulation study.

The first bootstrap method is based on Wu (1986). We first find the predictors of the random effects (v_i) and the random errors (e_{ij}), and then construct the bootstrap samples by combining the random effects and the random errors with weights drawn from an external distribution. The second bootstrap method applies the classical bootstrap procedure for the model parameters on the transformed approximately i.i.d. data, and then applies Solomon's parametric formulae to obtain the estimates of the three interested quantities. Our third bootstrap method involves data transformation and the classical bootstrap procedure. Similar to the second bootstrap method, we first apply the same data transformation, then draw the bootstrap samples on the transformed data, but then we apply the inverse transformation on the bootstrap samples, and finally compute the three interested quantities directly from transformed bootstrap samples.

To make the description clear, we use the following notations. Let:

$$N = \sum_{i=1}^{a} n,$$

$$\bar{y}_{i.} = \frac{\sum_{j=1}^{n} y_{ij}}{n},$$

$$\bar{y}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{N},$$

$$\bar{v}_{.} = \frac{\sum_{i=1}^{a} v_i}{a},$$

$$\bar{e}_{i.} = \frac{\sum_{j=1}^{n} e_{ij}}{n},$$

$$\bar{e}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} e_{ij}}{N},$$

$$Q_B = \sum_{i=1}^{a}\sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2,$$

$$Q_W = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2,$$

$$\Delta = \frac{\sigma_v^2}{\sigma_e^2},$$

$$\hat{\Delta} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2},$$

$$\alpha = 1 - \sqrt{\frac{1}{1+n\Delta}},$$

$$\hat{\alpha} = 1 - \sqrt{\frac{1}{1+n\hat{\Delta}}},$$

$O(n^{-1})$ represents terms that have the same or smaller order of magnitude as n^{-1}, in the sense that $\left|\frac{O(n^{-1})}{n^{-1}}\right|$ is bounded as $n \to \infty$.

$O\left((a \wedge n)^{-1}\right)$ denotes $O(a^{-1})$ or $O(n^{-1})$ or both of them.

E_M, Var_M and Cov_M denote the expectation, variance and covariance with respect to the model, and E_*, Var_*, and Cov_* denote the expectation, variance and covariance with respect to the bootstrap sample.

3.2 Bootstrap Method I

In this section, we applied Wu's external bootstrap procedure (Wu, 1986) to estimate the expected number of exceedances ($E(T)$), the variance of number of exceedances ($Var(T)$) and the probability of observing no exceedance ($Pr(T = 0)$) under one-way random effects model.

3.2.1 Predicting the Random Effects and Random Errors

Since in the model (2.1), we only have observed data y_{ij}, the overall mean μ, the random effect v_i and the random error e_{ij} are not observed, in order to apply Wu's procedure, we need to have the estimate of μ, and the predictions of v_i and e_{ij} first. The overall mean μ can be simply estimated by the overall sample mean, which is the ordinary least squared estimator:

$$\hat{\mu} = \bar{y}_{..} \tag{3.2}$$

For the random effect v_i and the random error e_{ij}, both are unobserved random variables, so we need to find their corresponding predictors $\hat{v}_i$ and $\hat{e}_{ij}$. There are many different sets of predictors, but, in order to mimic the behaviour of v_i and e_{ij}, we require that the predictors $\hat{v}_i$ and $\hat{e}_{ij}$ should have the same or asymptotically the same first and second order moments as v_i and e_{ij}, respectively, *i.e.*,we require that:

$$E(\hat{v}_i) = 0, \qquad E(\hat{e}_{ij}) = 0,$$

and,

$$Var(\hat{v}_i) \to \sigma_v^2, \qquad Var(\hat{e}_{ij}) \to \sigma_e^2, \qquad \text{as } a \to \infty \text{ and } n \to \infty.$$

According to the above requirements, we have four different sets of predictors. Our first set of predictors is to use the Best Linear Unbiased Predictor

(BLUP) of v_i, which is (assuming μ, σ_v^2 and σ_e^2 are known):

$$\hat{v}_i^A = \frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}(\bar{y}_{i\cdot} - \mu) \tag{3.3}$$

and,

$$\hat{e}_{ij}^A = y_{ij} - \mu - \hat{v}_i^A = y_{ij} - \frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}\bar{y}_{i\cdot} - \frac{\sigma_e^2}{n\sigma_v^2 + \sigma_e^2}\mu \tag{3.4}$$

It is easy to see that

$$E(\hat{v}_i^A) = \frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}(E\bar{y}_{i\cdot} - \mu) = 0,$$

and,

$$E(\hat{e}_{ij}^A) = E(y_{ij} - \mu - \hat{v}_i^A) = \mu - \mu - 0 = 0.$$

For the variances of $\hat{v}_i^A$ and $\hat{e}_{ij}^A$:

$$\begin{aligned}
Var(\hat{v}_i^A) &= Var\left(\frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}(\bar{y}_{i\cdot} - \mu)\right) \\
&= \frac{n^2\sigma_v^4}{(n\sigma_v^2 + \sigma_e^2)^2} Var(\bar{y}_{i\cdot}) \\
&= \frac{n^2\sigma_v^4}{(n\sigma_v^2 + \sigma_e^2)^2}\left(\sigma_v^2 + \frac{\sigma_e^2}{n}\right) \\
&= \frac{\sigma_v^4}{\sigma_v^2 + \frac{\sigma_e^2}{n}} \\
&\to \sigma_v^2, \qquad \text{as } n \to \infty
\end{aligned}$$

$$\begin{aligned}
Var(\hat{e}_{ij}^A) &= Var\left(y_{ij} - \frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}\bar{y}_{i\cdot} - \frac{\sigma_e^2}{n\sigma_v^2 + \sigma_e^2}\mu\right) \\
&= Var(y_{ij}) + \frac{n^2\sigma_v^4}{(n\sigma_v^2 + \sigma_e^2)^2} Var(\bar{y}_{i\cdot}) - 2\frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2} Cov(y_{ij}, \bar{y}_{i\cdot}) \\
&= \sigma_e^2 + \sigma_v^2 + \frac{n^2\sigma_v^4}{(n\sigma_v^2 + \sigma_e^2)^2}\left(\sigma_v^2 + \frac{\sigma_e^2}{n}\right) - 2\frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}\left(\sigma_v^2 + \frac{\sigma_e^2}{n}\right) \\
&= \sigma_e^2 - \sigma_v^2 + \frac{\sigma_v^4}{\sigma_v^2 + \frac{\sigma_e^2}{n}} \\
&\to \sigma_e^2, \qquad \text{as } n \to \infty
\end{aligned}$$

This shows that our first set of predictors $\hat{v}_i^A$ and $\hat{e}_{ij}^A$ has the same first order moment and asymptotically the same second order moment as the random effects v_i and the random effects e_{ij}, respectively.

But $\hat{v}_i^A$ and $\hat{e}_{ij}^A$ involve the unknown parameters μ, σ_v^2 and σ_e^2, we may replace them with their ANOVA estimators, which is described in Section (2.1).

Our second, third and fourth sets of predictors are constructed based on Harris & Burch (2003,2004). The second set is a simple, or naive predictors:

$$\hat{v}_i^B = \bar{y}_{i.} - \bar{y}_{..} \tag{3.5}$$

$$\hat{e}_{ij}^B = y_{ij} - \bar{y}_{i.} \tag{3.6}$$

It is easy to show that

$$E(\hat{v}_i^B) = E(\bar{y}_{i.} - \bar{y}_{..}) = \mu - \mu = 0$$

$$E(\hat{e}_{ij}^B) = E(y_{ij} - \bar{y}_{i.}) = \mu - \mu = 0$$

and

$$Var(\hat{v}_i^B) = \left(1 - \frac{1}{a}\right)\left(\frac{n\sigma_v^2 + \sigma_e^2}{n\sigma_v^2}\right)\sigma_v^2 \tag{3.7}$$

$$Var(\hat{e}_{ij}^B) = \left(1 - \frac{1}{n}\right)\sigma_e^2 \tag{3.8}$$

From the above equations, we can see that $\hat{v}_i^B$ and $\hat{e}_{ij}^B$ have the same expectation as v_i and e_{ij}, but their variances are not exactly equal to the variances of v_i and e_{ij}. But if $a \to \infty$ and $n \to \infty$, we have:

$$\lim_{\substack{a\to\infty \\ n\to\infty}} Var(\hat{v}_i^B) = \sigma_v^2$$

$$\lim_{n\to\infty} Var(\hat{e}_{ij}^B) = \sigma_e^2$$

So, their variances are asymptotic the same as v_i and e_{ij}.

From (3.7) and (3.8), we have:

$$Var\left[\sqrt{\frac{1}{\left(1-\frac{1}{a}\right)\left(\frac{n\sigma_v^2+\sigma_e^2}{n\sigma_v^2}\right)}}\hat{v}_i^B\right] = \sigma_v^2$$

$$Var\left[\sqrt{\frac{1}{1-\frac{1}{n}}}\hat{e}_{ij}^{B}\right] = \sigma_e^2$$

Thus we may construct a third set of predictors as (assuming σ_v^2 and σ_e^2 are known):

$$\hat{v}_i^C = \sqrt{\frac{1}{\left(1-\frac{1}{a}\right)\left(\frac{n\sigma_v^2+\sigma_e^2}{n\sigma_v^2}\right)}}\hat{v}_i^B = \sqrt{\frac{a}{a-1}\frac{n\sigma_v^2}{n\sigma_v^2+\sigma_e^2}}(\bar{y}_{i.}-\bar{y}_{..})$$

$$\hat{e}_{ij}^C = \sqrt{\frac{1}{1-\frac{1}{n}}}\hat{e}_{ij}^B = \sqrt{\frac{n}{n-1}}(y_{ij}-\bar{y}_{i.}).$$

Then, it is obvious that assuming σ_v^2 and σ_e^2 are known, $\hat{v}_i^C$ and $\hat{e}_{ij}^C$ have the same first and second order moments as v_i and e_{ij}, respectively. Again, we could replace σ_v^2 and σ_e^2 with their ANOVA estimators if these two parameters are unknown. But instead of replacing σ_v^2 and σ_e^2 separately with their estimators, we notice that under normal distribution assumption, we have:

$$E\left[\frac{(a-3)Q_W}{a(n-1)Q_B}\right] = \frac{\sigma_e^2}{n\sigma_v^2+\sigma_e^2}$$

so

$$E\left[1-\frac{(a-3)Q_W}{a(n-1)Q_B}\right] = \frac{n\sigma_v^2}{n\sigma_v^2+\sigma_e^2} \tag{3.9}$$

thus the term $\frac{n\sigma_v^2}{n\sigma_v^2+\sigma_e^2}$ in $\hat{v}_i^C$ can be unbiasedly estimated by $1-\frac{(a-3)Q_W}{a(n-1)Q_B}$, so we replace $\frac{n\sigma_v^2}{n\sigma_v^2+\sigma_e^2}$ in $\hat{v}_i^C$ with $1-\frac{(a-3)Q_W}{a(n-1)Q_B}$, then finally, our third set of predictors are:

$$\hat{v}_i^C = \sqrt{\max\left\{0, \frac{a}{a-1}\left[1-\frac{(a-3)Q_W}{a(n-1)Q_B}\right]\right\}}(\bar{y}_{i.}-\bar{y}_{..}) \tag{3.10}$$

$$\hat{e}_{ij}^C = \sqrt{\frac{n}{n-1}}(y_{ij}-\bar{y}_{i.}) \tag{3.11}$$

Finally, if we look at the above predictor of v_i, we see that Q_W and Q_B are correlated with $(\bar{y}_{i.}-\bar{y}_{..})$. Thus $E(\hat{v}_i^C)$ and $Var(\hat{v}_i^C)$ may not equal to $E(v_i)$ and

$Var(v_i)$. We then construct a fourth set of predictors for v_i using Jackknife method: Q_W and Q_B are computed by excluding the i^{th} group, denoted as $Q_{W(-i)}$ and $Q_{B(-i)}$. Then, similar to (3.9), we have:

$$E\left[1 - \frac{(a-4)Q_{W(-i)}}{(a-1)(n-1)Q_{B(-i)}}\right] = \frac{n\sigma_v^2}{n\sigma_v^2 + \sigma_e^2}$$

and, the fourth set of predictors is:

$$\hat{v}_i^D = \sqrt{\max\left\{0, \frac{a}{a-1}\left[1 - \frac{(a-4)Q_{W(-i)}}{(a-1)(n-1)Q_{B(-i)}}\right]\right\}}(\bar{y}_{i.} - \bar{y}_{..}) \qquad (3.12)$$

$$\hat{e}_{ij}^D = \sqrt{\frac{n}{n-1}}(y_{ij} - \bar{y}_{i.}) \qquad (3.13)$$

These four sets of predictors all have the required first and second order moments properties, and can be used in our first bootstrap method. Our simulation study shows that the Jackknife predictor works better than other predictors, so we will use the Jackknife predictor in this thesis.

3.2.2 Description of Bootstrap Method I

After obtaining the predictors of v_i and e_{ij}, we can then apply Wu's bootstrap procedure. Also considering Liu's suggestion (Liu, 1988) to obtain the same second order asymptotic properties as of the classical bootstrap for i.i.d. models, we obtain bootstrap weights (t_i^*) from $t_i^* \sim \Gamma(4,2) - 2$ distribution, where $\Gamma(4,2)$ denotes the Gamma distribution with shape parameter 4 and scale parameter 2, whose density function is: $f(x) = \frac{2^4}{\Gamma(4)}x^{4-1}e^{-2x}$. By choosing the distribution of the bootstrap weights to be $\Gamma(4,2) - 2$, we have the moment requirement suggested by Liu (1988): $E(t_i^*) = 0$, $E(t_i^{*2}) = 1$ and $E(t_i^{*3}) = 1$. The procedure is described in the following steps (Thach, 1998):

Step 1: Compute the predictors of v_i and e_{ij} using one of the above four predictors, let the predictors be $\hat{v}_i$ and $\hat{e}_{ij}$.

Step 2: Independently generate random numbers $\{t_i^*\}$ and $\{t_{ij}^*\}$:

$$t_i^* \sim \Gamma(4,2) - 2, \qquad i = 1, \ldots, a$$

and

$$t_{ij}^* \sim \Gamma(4,2) - 2, \qquad i = 1, \ldots, a, \qquad j = 1, \ldots, n$$

Step 3: Construct the bootstrap sample:

$$y_{ij}^* = \hat{\mu} + t_i^* \hat{v}_i + t_{ij}^* \hat{e}_{ij}$$

Step 4: Compute the following bootstrap estimators from the bootstrap sample:

$$E(\hat{T}^*) = \frac{1}{a} \sum_{i=1}^{a} \hat{T}_i^*$$

$$Var(\hat{T}^*) = \frac{1}{a-1} \sum_{i=1}^{a} \left(\hat{T}_i^* - E(\hat{T}^*) \right)^2$$

$$Pr(\hat{T}^* = 0) = \frac{\sum_{i=1}^{a} \prod_{j=1}^{n} I_{(y_{ij}^* \leq h)}}{a}$$

Where

$$\hat{T}_i^* = \sum_{j=1}^{n} I_{(y_{ij}^* > h)}$$

and

$$I_{(y_{ij}^* > h)} = \begin{cases} 1 & y_{ij}^* > h \\ 0 & Otherwise \end{cases}$$

Step 5: Repeat steps (2) to (4) B times, where B is a large number, and obtain B estimates for the three corresponding quantities, namely: $E_{(1)}(\hat{T}^*), \ldots, E_{(B)}(\hat{T}^*)$, $Var_{(1)}(\hat{T}^*), \ldots, Var_{(B)}(\hat{T}^*)$, and $Pr_{(1)}(\hat{T}^* = 0), \ldots, Pr_{(B)}(\hat{T}^* = 0)$. Then, the bootstrap estimators are:

$$E(T) = \frac{1}{B} \sum_{b=1}^{B} E_{(b)}(\hat{T}^*)$$

$$Var(T) = \frac{1}{B}\sum_{b=1}^{B} Var_{(b)}(\hat{T}^*)$$

and

$$Pr(T = 0) = \frac{1}{B}\sum_{b=1}^{B} Pr_{(b)}(\hat{T}^* = 0)$$

3.2.3 Consistency of the Bootstrap Estimators

To show the bootstrap estimators of $E(T)$, $Var(T)$ and $Pr(T = 0)$ are consistent, it is sufficient to show that the bootstrap sample y_{ij}^* and the observed data y_{ij} have asymptotically the same first and second order moments. We show this result is true.

Suppose:

(1) The model (2.1) is valid;

(2) The predicted random effects $\hat{v}_i$ and random errors $\hat{e}_{ij}$ have asymptotically the same first and second order moments with v_i and e_{ij}, respectively, *i.e.*,

$$E(\hat{v}_i) = O\left((a \wedge n)^{-1}\right) \tag{3.14}$$

$$E(\hat{e}_{ij}) = O\left((a \wedge n)^{-1}\right) \tag{3.15}$$

and

$$E(\hat{v}_i^2) = E(v_i^2) + O\left((a \wedge n)^{-1}\right) = \sigma_v^2 + O\left((a \wedge n)^{-1}\right) \tag{3.16}$$

$$E(\hat{e}_{ij}^2) = E(e_{ij}^2) + O\left((a \wedge n)^{-1}\right) = \sigma_e^2 + O\left((a \wedge n)^{-1}\right) \tag{3.17}$$

Then, (A)

$$E_M\left[E_*(y_{ij}^*)\right] = \mu \tag{3.18}$$

(B)

$$E_M\left[Cov_*(y_{ij}^*, y_{i'j'}^*\right] = \begin{cases} \sigma_v^2 + \sigma_e^2 + O\left((a \wedge n)^{-1}\right) & \text{for } i = i' \text{ and } j = j' \\ \sigma_v^2 + O\left((a \wedge n)^{-1}\right) & \text{for } i = i' \text{ and } j \neq j' \\ 0 & \text{otherwise} \end{cases} \tag{3.19}$$

where E_* and E_M represent expectation with respect to the distributions induced by bootstrap sampling and the model, respectively, Var_* represents variance with respect to the bootstrap samples, and $Cov_*(y^*_{ij}, y^*_{i'j'}) = E_*(y^*_{ij} - \hat{\mu})(y^*_{i'j'} - \hat{\mu})$.

Proof. **For (A):**

Under model (2.1), $\hat{\mu}$ is unbiased for μ, *i.e.*,$E_M(\hat{\mu}) = \mu$.

$$\begin{aligned} E_M\left[E_*(y^*_{ij})\right] &= E_M\left[E_*(\hat{\mu} + t^*_i\hat{v}_i + t^*_{ij}\hat{e}_{ij})\right] \\ &= E_M(\hat{\mu} + 0\hat{v}_i + 0\hat{e}_{ij}) \\ &= \mu \end{aligned} \tag{3.20}$$

For (B):

(1) If $i = i'$ and $j = j'$:

$$\begin{aligned} E_M\left[Var_*(y^*_{ij})\right] &= E_M\left[Var_*(\hat{\mu} + t^*_i\hat{v}_i + t^*_{ij}\hat{e}_{ij})\right] \\ &= E_M\left[\hat{v}_i^2 Var_*(t^*_i) + \hat{e}^2_{ij}Var_*(t_{ij})\right] \\ &= E_M(\hat{v}_i^2 + \hat{e}^2_{ij}) \\ &= E_M(\hat{v}_i^2) + E_M(\hat{e}^2_{ij}) \\ &= \sigma_v^2 + \sigma_e^2 + O\left((a \wedge n)^{-1}\right) \end{aligned} \tag{3.21}$$

(2) If $i = i'$ and $j \neq j'$:

$$\begin{aligned} & E_M \left[Cov_*(y_{ij}^*, y_{ij'}^*)\right] \\ = \; & E_M \left[Cov_*(\hat{\mu} + t_i^* \hat{v}_i + t_{ij}^* \hat{e}_{ij}, \hat{\mu} + t_i^* \hat{v}_i + t_{ij'}^* \hat{e}_{ij'})\right] \\ = \; & E_M \left[\hat{v}_i^2 Var_*(t_i^*)\right] \\ = \; & E_M(\hat{v}_i^2) \\ = \; & \sigma_v^2 + O\left((a \wedge n)^{-1}\right) \end{aligned} \tag{3.22}$$

Since t_i^*, t_{ij}^* and $t_{ij'}^*$ are independent.

(3) If $i \neq i'$ and $j \neq j'$"

$$\begin{aligned} & E_M \left[Cov_*(y_{ij}^*, y_{ij'}^*)\right] \\ = \; & E_M \left[Cov_*(\hat{\mu} + t_i^* \hat{v}_i + t_{ij}^* \hat{e}_{ij}, \hat{\mu} + t_i'^* \hat{v}_i' + t_{i'j'}^* \hat{e}_{i'j'})\right] \\ = \; & 0 \end{aligned} \tag{3.23}$$

Since t_i^*, $t_i'^*$, t_{ij}^* and $t_{ij'}^*$ are independent.

3.3 Bootstrap Method II

The second bootstrap method can be viewed as a semi-parametric method. Assume the model is correct, it first estimates the model parameters from the observed data. Then it uses the estimated parameters to construct a data transformation, trying to transform the correlated data to uncorrelated and identically distributed data. It then applies the classical bootstrap method

on the transformed data, and obtains the bootstrap estimates for the model parameters. Finally, it uses Solomon's parametric formulae (2.17), (2.18), and (2.19) to compute the estimates of the interested quantities. Here bootstrap samples depend on the data transformation, which relies on the estimates of the model parameters, and on the final step, the parametric formulae are also used, thus it is a combination of the parametric method and the bootstrap method.

3.3.1 Model Transformation

If the model (2.1) is correct, and if we know the two variance components σ_v^2 and σ_e^2, then, following Fuller & Battese (1973), define:

$$z_{ij} = y_{ij} - \alpha \bar{y}_{i.} \tag{3.24}$$

Then it is easy to see that:

$$E(z_{ij}) = E(y_{ij} - \alpha \bar{y}_{i.}) = \mu - \alpha\mu = (1-\alpha)\mu \tag{3.25}$$

and

$$Cov(z_{ij}, z_{i'j'}) = \begin{cases} \sigma_e^2 & \text{for } i = i' \text{ and } j = j' \\ 0 & \text{otherwise} \end{cases} \tag{3.26}$$

That is, z_{ij} are uncorrelated and identically distributed, with mean $\mu_z = (1-\alpha)\mu$ and variance $\sigma_z^2 = \sigma_e^2$. Under normal assumption, z_{ij} are i.i.d., and we can apply the classical bootstrap procedure on z_{ij}, to obtain the bootstrap

estimator for the mean and variance, μ_z^* and σ_z^{*2}, respectively. If we have the estimates of μ_z and σ_z^2, then, we can have the estimates of μ, σ_v^2 and σ_e^2 by:

$$\hat{\mu} = \frac{\hat{\mu}_z}{1-\alpha} \tag{3.27}$$

$$\hat{\sigma}_v^2 = \Delta\hat{\sigma}_z^2 \tag{3.28}$$

$$\hat{\sigma}_e^2 = \hat{\sigma}_z^2 \tag{3.29}$$

Then using the formulae (2.17), (2.18), and (2.19), we can get the estimates for the three interested quantities $E(T)$, $Var(T)$ and $Pr(T=0)$.

But the transformation (3.24) requires knowing both σ_v^2 and σ_e^2. Hence, we need to estimate these two values from the observed data y_{ij}. This can be done by using the ANOVA estimators:

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{a(n-1)} \tag{3.30}$$

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2}{a-1} - \frac{\hat{\sigma}_e^2}{n} \tag{3.31}$$

3.3.2 Description of Bootstrap Method II

The second bootstrap method is now described as below:

Step 1: Compute the ANOVA estimators for σ_e^2 and σ_v^2 from the data

$$\hat{\sigma}_e^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2}{a(n-1)}$$

and

$$\hat{\sigma}_v^2 = \frac{\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1} - \frac{\hat{\sigma}_e^2}{n}$$

Step 2: If $\hat{\sigma}_v^2 < 0$, then let

$$\hat{\sigma}_v^2 = 0, \qquad \hat{\sigma}_e^2 = \frac{\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2}{an-1}$$

Step 3: Compute $\hat{\alpha}$ and $\hat{\Delta}$:

$$\hat{\alpha} = 1 - \sqrt{\frac{1}{1+n\hat{\Delta}}}$$

$$\hat{\Delta} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2}$$

Step 4: Do the following transformation:

$$z_{ij} = y_{ij} - \hat{\alpha}\bar{y}_{i.}$$

Step 5: Compute:

$$\bar{z}_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} z_{ij}}{an}$$

Step 6: Generate $a \times n$ pairs of random number: i_k^* and j_k^*, $k = 1, \ldots, a \times n$, where i_k^* is an integer number from uniform distribution $U(1, a)$, and j_k^* is an integer number from uniform distribution $U(1, n)$. Then, a bootstrap sample is defined as:

$$z_k^* = z_{i_k^* j_k^*}, \qquad k = 1, \ldots, a \times n$$

Step 7: From z_k^*, we compute $\hat{\mu}_z^*$, and $\hat{\sigma}_z^{*2}$:

$$\hat{\mu}_z^* = \frac{\sum_{k=1}^{an} z_k^*}{an}, \qquad \hat{\sigma}_z^{*2} = \frac{\sum_{k=1}^{an}(z_k^* - \bar{z}_{..})^2}{an-1}$$

Step 8: Compute $\hat{\mu}^*$, $\hat{\sigma}_e^{*2}$, and $\hat{\sigma}_v^{*2}$:

$$\hat{\mu}^* = \frac{\hat{\mu}_z^*}{1-\hat{\alpha}}, \qquad \hat{\sigma}_e^{*2} = \hat{\sigma}_z^{*2}, \qquad \hat{\sigma}_v^{*2} = \hat{\Delta}\hat{\sigma}_e^{*2}$$

Step 9: Plug in $\hat{\mu}^*$, $\hat{\sigma}_e^{*2}$, and $\hat{\sigma}_v^{*2}$ in Solomon's formula:

$$E^*(T) = n\Phi\left(\frac{\hat{\mu}^* - h}{\hat{\sigma}^*}\right)$$

$$Var^*(T) = n\Phi\left(\frac{\hat{\mu}^* - h}{\hat{\sigma}^*}\right)\left[1 - \Phi\left(\frac{\hat{\mu}^* - h}{\hat{\sigma}^*}\right)\right][1 + (n-1)\hat{\rho}_I^*]$$

$$Pr^*(T=0) = \int_{-\infty}^{\infty} \phi(x)\left\{\Phi\left(\frac{\hat{\gamma}^* - x}{\hat{\tau}^*}\right)\right\}^n dx$$

where

$$\hat{\rho}_I^* = \frac{\Phi_2\left(\frac{\hat{\mu}^*-h}{\hat{\sigma}^*}, \frac{\hat{\mu}^*-h}{\hat{\sigma}^*}, \hat{\rho}^* = \frac{\hat{\sigma}_v^{*2}}{\hat{\sigma}^{*2}}\right) - \Phi^2\left(\frac{\hat{\mu}^*-h}{\hat{\sigma}^*}\right)}{\Phi\left(\frac{\hat{\mu}^*-h}{\hat{\sigma}^*}\right)\left[1 - \Phi\left(\frac{\hat{\mu}^*-h}{\hat{\sigma}^*}\right)\right]}$$

and

$$\hat{\sigma}^{*2} = \hat{\sigma}_v^{*2} + \hat{\sigma}_e^{*2}, \qquad \hat{\gamma}^* = \frac{h - \hat{\mu}^*}{\hat{\sigma}_v^*}, \qquad \hat{\tau}^* = \frac{\hat{\sigma}_e^*}{\hat{\sigma}_v^*}$$

Step 10: Repeat Steps (6)-(9) B times, where B is a large number, then we have B estimates for the 3 corresponding quantities, namely: $E_{(1)}^*(T), \ldots, E_{(B)}^*(T)$, $Var_{(1)}^*(T), \ldots, Var_{(B)}^*(T)$, and $Pr_{(1)}^*(T=0), \ldots, Pr_{(B)}^*(T=0)$. Then, the corresponding estimates from bootstrap method II are:

$$E^*(T) = \frac{1}{B}\sum_{b=1}^{B} E_{(b)}^*(T)$$

$$Var^*(T) = \frac{1}{B}\sum_{b=1}^{B} Var^*_{(b)}(T)$$

and

$$Pr^*(T=0) = \frac{1}{B}\sum_{b=1}^{B} Pr^*_{(b)}(T=0)$$

3.3.3 Justification of the Bootstrap Method II

We prove that the bootstrap estimators for $E(T)$, $V(T)$ and $Pr(T=0)$ are consistent. This can be done by first proving that the bootstrap estimators $\hat{\mu}^*$, $\hat{\sigma}_v^{*2}$ and $\hat{\sigma}_e^{*2}$ are consistent for μ, σ_v^2, and σ_e^2, respectively. Since from (2.17), (2.18) and (2.19), we see that $E(T)$, $V(T)$ and $Pr(T=0)$ are all smooth functions of μ, σ_v^2 and σ_e^2, then the required results follow.

Note that under model (2.1), the estimators $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ are consistent for σ_v^2 and σ_e^2, respectively, *i.e.*,$E_M(\hat{\sigma}_v^2) = \sigma_v^2 + O(a^{-1})$, $E_M(\hat{\sigma}_e^2) = \sigma_e^2 + O(a^{-1})$.

Since $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ are consistent for σ_v^2 and σ_e^2, and $\hat{\alpha}(= 1 - \sqrt{\frac{1}{1+n\hat{\Delta}}}$, $\hat{\Delta} = \frac{\hat{\sigma}_v^2}{\hat{\sigma}_e^2})$ is smooth function of $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$, then, it follows that $\hat{\alpha}$ is consistent estimator of α, *i.e.*,$E_M(\hat{\alpha}) = \alpha + O(a^{-1})$. We then show the following conclusions:

(A) $E_M E_*(\hat{\mu}^*) = \mu + O(a^{-1})$;

(B) $E_M E_*(\hat{\sigma}_e^{*2}) = \sigma_e^2 + O(a^{-1})$;

(C) $E_M E_*(\hat{\sigma}_v^{*2}) = \sigma_v^2 + O(a^{-1})$.

Proof of (A):

(1) Since z_k^* is taken from $\{z_{ij} : i = 1, 2, \ldots, a; j = 1, 2, \ldots, n\}$ with probability $\frac{1}{an}$, then,

$$\begin{aligned} E_*(z_k^*) &= \sum_{i=1}^{a}\sum_{j=1}^{n} z_{ij}\frac{1}{an} = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij} - \hat{\alpha}\bar{y}_{i.}) \\ &= \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}\left[(1-\hat{\alpha})\mu + (1-\hat{\alpha})v_i + e_{ij} - \hat{\alpha}\bar{e}_{i.}\right] \end{aligned}$$

so,

$$\begin{aligned} E_M E_*(z_k^*) &= \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} E_M\left[(1-\hat{\alpha})\mu + (1-\hat{\alpha})v_i + e_{ij} - \hat{\alpha}\bar{e}_{i.}\right] \\ &= \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}\left[(1-\alpha)\mu + O(a^{-1})\right] \\ &= (1-\alpha)\mu + O(a^{-1}) \end{aligned}$$

thus,

$$\begin{aligned} E_M E_*(\hat{\mu}^*) &= E_M E_*\left[\frac{\hat{\mu}_z^*}{1-\hat{\alpha}}\right] = E_M E_*\left[\frac{\frac{1}{an}\sum_{k=1}^{an} z_k^*}{1-\hat{\alpha}}\right] \\ &= E_M \frac{1}{(1-\hat{\alpha})an}\sum_{k=1}^{an} E_*(z_k^*) \\ &= \frac{1}{(1-\alpha)an}\sum_{k=1}^{an}\left[(1-\alpha)\mu + O(a^{-1})\right] \\ &= \mu + O(a^{-1}) \end{aligned} \tag{3.32}$$

(A) is proved.

Proof of (B): we prove the conclusion of (B) in several steps:

STEP 1:

Since z_k^* is taken from $\{z_{ij} : i = 1, 2, \ldots, a; j = 1, 2, \ldots, n\}$ with probability

$\frac{1}{an}$, then

$$\begin{aligned}
E_* \quad & (z_k^* - \bar{z}_{..})^2 \\
= \quad & \sum_{i=1}^{a}\sum_{j=1}^{n}(z_{ij} - \bar{z}_{..})^2 \frac{1}{an} \\
= \quad & \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}[(y_{ij} - \hat{\alpha}\bar{y}_{i.}) - (\bar{y}_{..} - \hat{\alpha}\bar{y}_{..})]^2 \\
= \quad & \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}\{[(\mu + v_i + e_{ij}) - \hat{\alpha}(\mu + v_i + \bar{e}_{i.})] \\
- \quad & [(1-\hat{\alpha})(\mu + \bar{v}_{.} + \bar{e}_{..})]\}^2 \\
= \quad & \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}[(1-\hat{\alpha})(v_i - \bar{v}_{.}) + e_{ij} - \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}]^2
\end{aligned}$$

STEP 2: Because

$$\begin{aligned}
E_M \quad & [(1-\hat{\alpha})(v_i - \bar{v}_{.}) + e_{ij} - \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}] \\
= \quad & (1-\alpha)[E_M(v_i) - E_M(\bar{v}_{.})] + E_M(e_{ij}) - \alpha E_M(\bar{e}_{i.}) \\
- \quad & (1-\alpha)E_M(\bar{e}_{..}) + O(a^{-1}) \\
= \quad & O(a^{-1})
\end{aligned}$$

we have:

$$\begin{aligned}
E_M \quad & [(1-\hat{\alpha})(v_i - \bar{v}_{.}) + e_{ij} - \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}]^2 \\
\approx \quad & Var_M[(1-\hat{\alpha})(v_i - \bar{v}_{.}) + e_{ij} - \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}] \\
= \quad & (1-\alpha)^2 Var_M(v_i - \bar{v}_{.}) + Var_M[e_{ij} - \alpha\bar{e}_{i.} - (1-\alpha)\bar{e}_{..}] \\
+ \quad & O(a^{-1})
\end{aligned}$$

where

$$\begin{aligned} Var_M(v_i - \bar{v}_.) &= Var_M(v_i) + Var_M(\bar{v}_.) - 2Cov_M(v_i, \bar{v}_.) \\ &= \sigma_v^2 + \frac{\sigma_v^2}{a} - 2\frac{\sigma_v^2}{a} = \frac{a-1}{a}\sigma_v^2 \end{aligned}$$

and

$$\begin{aligned} Var_M & [e_{ij} - \alpha\bar{e}_{i.} - (1-\alpha)\bar{e}_{..}] \\ = & \; Var_M(e_{ij}) + \alpha^2 Var_M(\bar{e}_{i.}) + (1-\alpha)^2 Var_M(\bar{e}_{..}) \\ - & \; 2\alpha Cov_M(e_{ij}, \bar{e}_{i.}) - 2(1-\alpha)Cov_M(e_{ij}, \bar{e}_{..}) \\ + & \; 2\alpha(1-\alpha)Cov_M(\bar{e}_{i.}, \bar{e}_{..}) \\ = & \; \sigma_e^2 + \alpha^2\frac{\sigma_e^2}{n} + (1-\alpha)^2\frac{\sigma_e^2}{an} - 2\alpha\frac{\sigma_e^2}{n} \\ - & \; 2(1-\alpha)\frac{\sigma_e^2}{an} + 2\alpha(1-\alpha)\frac{\sigma_e^2}{an} \\ = & \; \sigma_e^2 - \frac{\sigma_e^2}{n} + (1-\alpha)^2\frac{\sigma_e^2}{n} - (1-\alpha)^2\frac{\sigma_e^2}{an} \\ = & \; \sigma_e^2 - \frac{\sigma_e^2}{n} + (1-\alpha)^2\frac{a-1}{a}\frac{\sigma_e^2}{n} \end{aligned}$$

so,

$$\begin{aligned} Var_M & [(1-\hat{\alpha})(v_i - \bar{v}_.) + e_{ij} - \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}] \\ = & \; \frac{\sigma_e^2}{\sigma_e^2 + n\sigma_v^2}\frac{a-1}{a}\sigma_v^2 + \sigma_e^2 - \frac{\sigma_e^2}{n} + (1-\alpha)^2\frac{a-1}{a}\frac{\sigma_e^2}{n} \\ = & \; \frac{a-1}{a}(1-\alpha)^2\left(\sigma_v^2 + \frac{\sigma_e^2}{n}\right) + \sigma_e^2 - \frac{\sigma_e^2}{n} \\ = & \; \frac{a-1}{a}\frac{\sigma_e^2}{n\sigma_v^2 + \sigma_e^2}\frac{n\sigma_v^2 + \sigma_e^2}{n} + \sigma_e^2 - \frac{\sigma_e^2}{n} \\ = & \; \frac{a-1}{a}\frac{\sigma_e^2}{n} - \frac{\sigma_e^2}{n} + \sigma_e^2 \\ = & \; \frac{an-1}{an}\sigma_e^2 \end{aligned}$$

and finally, we have

$$\begin{aligned} E_M E_*(z_k^* - \bar{z}_{..})^2 &= \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} E_M[(1-\hat{\alpha})(v_i - \bar{v}_{.}) + e_{ij} \\ &- \hat{\alpha}\bar{e}_{i.} - (1-\hat{\alpha})\bar{e}_{..}]^2 \\ &= \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n}\left[\frac{an-1}{an}\sigma_e^2 + O(a^{-1})\right] \\ &= \frac{an-1}{an}\sigma_e^2 + O(a^{-1}) \end{aligned}$$

STEP 3: Finally, for the bootstrap estimator $\hat{\sigma}_z^{*2}$:

$$\begin{aligned} E_M E_*\left(\hat{\sigma}_z^{*2}\right) &= E_M E_*\left[\frac{\sum_{k=1}^{an}(z_k^* - \bar{z}_{..})^2}{an-1}\right] \\ &= \frac{1}{an-1}\sum_{k=1}^{an} E_M E_*(z_k^* - \bar{z}_{..})^2 \\ &= \frac{1}{an-1}\sum_{k=1}^{an}\left[\frac{an-1}{an}\sigma_e^2 + O(a^{-1})\right] \\ &= \sigma_e^2 + O(a^{-1}) \end{aligned} \tag{3.33}$$

This proves (B).

Proof of (C):

Since $\hat{\sigma}_v^{*2} = \hat{\Delta}\hat{\sigma}_z^{*2}$, $\hat{\Delta}$ is consistent for Δ, $\hat{\sigma}_z^{*2}$ is consistent for σ_e^2 from conclusion (B), it follows that $\hat{\sigma}_v^{*2}$ is consistent for σ_v^2. Thus (C) is proved.

From the conclusions (A), (B) and (C), and noting that $E(T)$, $V(T)$ and $Pr(T = 0)$ are smooth functions of μ, σ_e^2 and σ_v^2, then, the bootstrap estimators, $E^*(T)$, $Var^*(T)$ and $Pr^*(T = 0)$, constructed by replacing σ_e^2 and σ_v^2 with their bootstrap estimators, are consistent for $E(T)$, $V(T)$ and $Pr(T = 0)$, respectively.

3.4 Bootstrap Method III

In the second bootstrap method, after having the bootstrap estimates of the two variance components from the i.i.d. normal $\{z_{ij}\}$, we still have to use Solomon's parametric formulae to compute the required quantities, which are derived under normal assumption, thus the robustness of the second bootstrap method maybe weak. It is desirable that the estimates of $E(T)$, $V(T)$ and $Pr(T = 0)$ can be computed directly from the bootstrap samples, thus we propose the third bootstrap method.

3.4.1 Description of Bootstrap Method III

The third bootstrap method is described as below:

Steps 1-4: Exactly the same as steps 1-4 in second bootstrap method.

Step 5: Generate an pairs of random numbers i_k^* and j_k^* in the same way as Step 6 in the second bootstrap method. Then arrange the bootstrap sample z_{ij}^* in the following way:

$$z_{11}^* = z_{i_1^* j_1^*}, \quad \ldots, \quad z_{1n}^* = z_{i_n^* j_n^*}$$

$$\ldots\ldots$$

$$z_{a1}^* = z_{i_{n(a-1)+1}^* j_{n(a-1)+1}^*}, \quad \ldots, \quad z_{an}^* = z_{i_{na}^* j_{na}^*}$$

Step 6: Do the following transformation:

$$y_{ij}^{*} = z_{ij}^{*} + \frac{\hat{\alpha}}{1-\hat{\alpha}}\bar{z}_{i.}^{*}$$

Step 7: Compute the following quantities:

$$E(\hat{T}^{*}) = \frac{1}{a}\sum_{i=1}^{a}\hat{T}_{i}^{*}$$

$$Var(\hat{T}^{*}) = \frac{1}{a-1}\sum_{i=1}^{a}\left(\hat{T}_{i}^{*} - E(\hat{T}^{*})\right)^{2}$$

$$Pr(\hat{T}^{*}=0) = \frac{\sum_{i=1}^{a}\prod_{j=1}^{n} I_{(y_{ij}^{*}\leq h)}}{a}$$

Where

$$\hat{T}_{i}^{*} = \sum_{j=1}^{n} I_{(y_{ij}^{*}>h)}$$

and

$$I_{(y_{ij}^{*}>h)} = \begin{cases} 1 & y_{ij}^{*} > h \\ 0 & \text{Otherwise} \end{cases}$$

Step 8: Repeat steps (5) to (7) B times, where B is a large number, and obtain B estimates for the three corresponding quantities, namely: $E_{(1)}(\hat{T}^{*}), \ldots, E_{(B)}(\hat{T}^{*})$, $Var_{(1)}(\hat{T}^{*}), \ldots, Var_{(B)}(\hat{T}^{*})$, and $Pr_{(1)}(\hat{T}^{*}=0), \ldots, Pr_{(B)}(\hat{T}^{*}=0)$. Then, the bootstrap estimators are:

$$E(T) = \frac{1}{B}\sum_{b=1}^{B} E_{(b)}(\hat{T}^{*})$$

$$Var(T) = \frac{1}{B}\sum_{b=1}^{B} Var_{(b)}(\hat{T}^{*})$$

and

$$Pr(T=0) = \frac{1}{B}\sum_{b=1}^{B} Pr_{(b)}(\hat{T}^{*}=0)$$

3.4.2 Consistency of the Bootstrap estimators

Now we will show that the bootstrap estimators from the above procedure are consistent for the three interested quantities. Similar to the first bootstrap method, it is sufficient to show that the bootstrap sample y_{ij}^* and the observed data y_{ij} asymptotically have the same first and second order moments, that is, we will show that:

(A)

$$E_M E_*(y_{ij}^*) = \mu + O\left(a^{-1}\right) \tag{3.34}$$

and (B)

$$E_M\left[Cov_*(y_{ij}^*, y_{i'j'}^*)\right] = \begin{cases} \sigma_v^2 + \sigma_e^2 + O\left(a^{-1}\right) & \text{for } i = i' \text{ and } j = j' \\ \sigma_v^2 + O\left(a^{-1}\right) & \text{for } i = i' \text{ and } j \neq j' \\ O\left(a^{-1}\right) & \text{otherwise} \end{cases} \tag{3.35}$$

Proof. Since $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ are consistent for σ_v^2 and σ_e^2, respectively, then $\hat{\alpha}$ is consistent for α, then we have:

$$z_{ij} = y_{ij} - \hat{\alpha}\bar{y}_{i\cdot} \xrightarrow{L} y_{ij} - \alpha\bar{y}_{i\cdot} \sim N\left((1-\alpha)\mu, \sigma_e^2\right)$$

where $\xrightarrow{L}$ denotes convergence in law. This shows that z_{ij} are asymptotically i.i.d, which we have:

$$E_M(z_{ij}) = (1-\alpha)\mu + O(a^{-1}) \tag{3.36}$$

$$Cov_M(z_{ij}, z_{i'j'}) = \begin{cases} \sigma_e^2 + O(a^{-1}) & \text{for } i = i' \text{ and } j = j' \\ O(a^{-1}) & \text{otherwise} \end{cases} \tag{3.37}$$

Since z_{ij}^* is a bootstrap sample from z_{ij}, it is easy to show that (using the results of (3.36) and (3.37), and the same approach in Section 3.3):

$$E_M E_*(z_{ij}^*) = (1-\alpha)\mu + O(a^{-1}) \tag{3.38}$$

$$E_M Cov_*(z_{ij}^*, z_{i'j'}^*) = \begin{cases} \sigma_e^2 + O(a^{-1}) & \text{for } i = i' \text{ and } j = j' \\ O(a^{-1}) & \text{otherwise} \end{cases} \tag{3.39}$$

For (A):

$$\begin{aligned} E_M E_*(y_{ij}^*) &= E_M E_*(z_{ij}^*) + E_M\left(\frac{\hat{\alpha}}{1-\hat{\alpha}} E_* \bar{z}_{i.}^*\right) \\ &= (1-\alpha)\mu + \frac{\alpha}{1-\alpha}(1-\alpha)\mu + O(a^{-1}) \\ &= \mu + O(a^{-1}) \end{aligned} \tag{3.40}$$

This proves (3.34)

For (B) From (3.39), we have:

$$E_M Var_*(\bar{z}_{i.}^*) = \frac{\sigma_e^2}{n} + O\left(a^{-1}\right) \tag{3.41}$$

$$E_M Cov_*(\bar{z}_{i.}^*, \bar{z}_{i'.}^*) = O(a^{-1}) \qquad \text{for } i \neq i' \tag{3.42}$$

$$E_M Cov_*(z_{ij}^*, \bar{z}_{i'.}^*) = \begin{cases} \frac{\sigma_e^2}{n} + O(a^{-1}) & i = i' \\ O(a^{-1}) & i \neq i' \end{cases} \tag{3.43}$$

Then, (i) for $i = i'$ and $j = j'$, we have:

$$\begin{aligned}
E_M Var_*(y^*_{ij}) &= E_M Var_* \left(z^*_{ij} + \frac{\hat{\alpha}}{1-\hat{\alpha}} \bar{z}^*_{i.} \right) \\
&= E_M Var_*(z^*_{ij}) + 2E_M \left(\frac{\hat{\alpha}}{1-\hat{\alpha}} \right) Cov_*(z^*_{ij}, \bar{z}^*_{i.}) \\
&+ E_M \left(\frac{\hat{\alpha}}{1-\hat{\alpha}} \right)^2 Var_*(\bar{z}^*_{i.}) \\
&= \sigma_e^2 + 2\frac{\alpha}{1-\alpha}\frac{\sigma_e^2}{n} + \left(\frac{\alpha}{1-\alpha} \right)^2 \frac{\sigma_e^2}{n} + O\left(a^{-1}\right) \\
&= \sigma_e^2 + \frac{2\alpha - \alpha^2}{(1-\alpha)^2}\frac{\sigma_e^2}{n} + O\left(a^{-1}\right) \\
&= \sigma_e^2 + \Delta\sigma_e^2 + O\left(a^{-1}\right) \\
&= \sigma_e^2 + \sigma_v^2 + O\left(a^{-1}\right) \qquad (3.44)
\end{aligned}$$

since $\Delta = \frac{\sigma_v^2}{\sigma_e^2}$.

(ii) For $i = i'$ and $j \neq j'$:

$$\begin{aligned}
E_M \;& Cov_*(y^*_{ij}, y^*_{ij'}) = E_M Cov_*(z^*_{ij} + \frac{\hat{\alpha}}{1-\hat{\alpha}}\bar{z}^*_{i.}, z^*_{ij'} + \frac{\hat{\alpha}}{1-\hat{\alpha}}\bar{z}^*_{i.}) \\
&= E_M Cov_*(z^*_{ij}, z^*_{ij'}) + E_M \left[\frac{\hat{\alpha}}{1-\hat{\alpha}} Cov_*(z^*_{ij}, \bar{z}^*_{i.}) \right] \\
&+ E_M \left[\frac{\hat{\alpha}}{1-\hat{\alpha}} Cov_*(z^*_{ij'}, \bar{z}^*_{i.}) \right] + E_M Var_*(\bar{z}^*_{i.}) \\
&= 2\frac{\alpha}{1-\alpha}\frac{\sigma_e^2}{n} + \left(\frac{\alpha}{1-\alpha} \right)^2 \frac{\sigma_e^2}{n} + O(a^{-1}) \\
&= \Delta\sigma_e^2 + O(a^{-1}) \\
&= \sigma_v^2 + O(a^{-1}) \qquad (3.45)
\end{aligned}$$

And finally, (iii) for $i \neq i'$ and $j \neq j'$:

$$\begin{aligned} E_M \quad & Cov_*(y^*_{ij}, y^*_{i'j'}) \\ = \quad & E_M Cov_*(z^*_{ij} + \frac{\hat{\alpha}}{1-\hat{\alpha}}\bar{z}^*_{i.}, z^*_{i'j'} + \frac{\hat{\alpha}}{1-\hat{\alpha}}\bar{z}^*_{i'.}) \\ = \quad & E_M Cov_*(z^*_{ij}, z^*_{i'j'}) + E_M\left[\frac{\hat{\alpha}}{1-\hat{\alpha}} Cov_*(z^*_{ij}, \bar{z}^*_{i'.})\right] \\ + \quad & E_M\left[\frac{\hat{\alpha}}{1-\hat{\alpha}} Cov_*(z^*_{i'j'}, \bar{z}^*_{i.})\right] + E_M Cov_*(\bar{z}^*_{i.}, \bar{z}^*_{i'.}) \\ = \quad & O(a^{-1}) \end{aligned} \tag{3.46}$$

Combining (i), (ii) and (iii), we prove (3.35)

Chapter 4

Simulation Studies and Some Conclusions

The performances of the parametric method and the bootstrap methods developed in Chapter 2 and Chapter 3 are studied in this chapter. As an example, we gave an application of all these methods to a blood pressure data set at the end of this chapter.

4.1 Simulation Results

We use a simulation study to investigate the relative performances of the three bootstrap methods described in previous chapter and Solomon's parametric method. We assume that the model equation is true, that is: $y_{ij} = \mu + v_i + e_{ij}$, $i = 1, 2, \ldots, a$, $j = 1, 2, \ldots, n$, but the distribution of the random effects $\{v_i\}$

and the random errors $\{e_{ij}\}$ may not be normal. In this study, we consider the following four different distributional cases for $\{v_i\}$ and $\{e_{ij}\}$:

- CASE 1, Normal distribution:

$$v_i \sim N(0, 26.86),$$

$$e_{ij} \sim N(0, 52.02);$$

- CASE 2, Mixed normal distribution:

$$v_i \sim 0.1N(0, 26, 86) + 0.9N(0, 52.02),$$

$$e_{ij} \sim 0.1N(0, 26, 86) + 0.9N(0, 52.02);$$

- CASE 3, Cauchy distribution:

$$v_i \sim Cauchy(0, 1),$$

$$e_{ij} \sim Cauchy(0, 1);$$

- CASE 4, Double Exponential distribution:

$$v_i \sim DE(0, 26.86),$$

$$e_{ij} \sim DE(0, 52.02).$$

and we select the number of groups $a = 25$, number of subjects within each group $n = 16$, the grand mean $\mu = 91.70$, and the threshold $h = 95$. These

values (except for CASE 3) are computed from the actual blood pressure data set from the International Prospective Primary Prevention Study in Hypertension (IPPPSH). Because the variance for a Cauchy distribution does not exist, we use the standard Cauchy distribution with location parameter 0 and scale parameter 1 for CASE 3.

We generate $N = 1000$ sets of simulated random numbers $\{v_i^{(k)}\}$ and $\{e_{ij}^{(k)}\}$, where $i = 1, 2, \ldots, 25$, $j = 1, 2, \ldots, 16$, and $k = 1, 2, \ldots, N$, according to one of the four distributions specified above. Then, for the k^{th} set of simulated values of $\{v_i^{(k)}\}$ and $\{e_{ij}^{(k)}\}$, we construct the k^{th} set of simulated data $\{y_{ij}^{(k)}\}$ using equation $y_{ij}^{(k)} = \mu + v_i^{(k)} + e_{ij}^{(k)}$. We then apply all the methods developed in Chapters 2 and 3 to $\{y_{ij}^{(k)}\}$ (for the first bootstrap method, we use the Jackknife predictor of v_i and e_{ij}), and estimate all the three quantities from the k^{th} set of data, $\{y_{ij}^{(k)}\}$: expected number of exceedance $E^{(k)}(T)$, variance of number of exceedance $Var^{(k)}(T)$, and the probability of observing no exceedance $Pr^{(k)}(T = 0)$. Finally, the estimated values for the three quantities are the average of these three quantities over all sets of simulated data, respectively, *i.e.*,$E_{est}(T) = \frac{\sum_{k=1}^{N} E^{(k)}(T)}{N}$, $Var_{est}(T) = \frac{\sum_{k=1}^{N} Var^{(k)}(T)}{N}$, $Pr_{est}(T = 0) = \frac{\sum_{k=1}^{N} Pr^{(k)}(T=0)}{N}$.

From N sets of data $\{y_{ij}^{(k)}\}$, $k = 1, 2, \ldots, N$, we can compute the true values for $E(T)$, $Var(T)$ and $Pr(T = 0)$ in the following way:

For the k^{th} set of simulated data $\{y_{ij}^{(k)}\}$, first compute the number of exceedance in the i^{th} group:

$$T_i^k = \sum_{j=1}^{n} I_{\left(y_{ij}^{(k)} > h\right)} \tag{4.1}$$

Then compute:

Average number of exceedances: $E_{obs}^k(T) = \frac{\sum_{i=1}^{a} T_i^k}{a}$;

Variance of the number of exceedances: $Var_{obs}^k(T) = \frac{\sum_{i=1}^{a}(T_i^k - E_{obs}^k(T))^2}{a-1}$;

Probability of observing no exceedance: $Pr_{obs}^k(T=0) = \frac{\sum_{i=1}^{a}\prod_{j=1}^{n} I_{\left(y_{ij}^{(k)} \leq h\right)}}{a}$.

From N sets of simulated data $\{y_{ij}^{(k)}\}$, we get N values of $E_{obs}^k(T)$, $Var_{obs}^k(T)$ and $Pr_{obs}^k(T=0)$, $k = 1, 2, \ldots, N$, then the average of these N values are considered to be the true values of number of exceedances, variance of number of exceedances and probability of observing no exceedance, respectively, *i.e.*, $E_{true}(T) = \frac{\sum_{k=1}^{N} E_{obs}^k(T)}{N}$, $Var_{true}(T) = \frac{\sum_{k=1}^{N} Var_{obs}^k(T)}{N}$, $Pr_{true}(T=0) = \frac{\sum_{k=1}^{N} Pr_{obs}^k(T=0)}{N}$.

The performance of each method is measured by the relative bias, which is defined as below:

$$\text{Relative Bias of } E(T) = \left|\frac{E_{est}(T) - E_{true}(T)}{E_{true}(T)}\right| \times 100\% \tag{4.2}$$

$$\text{Relative Bias of } Var(T) = \left|\frac{Var_{est}(T) - Var_{true}(T)}{Var_{true}(T)}\right| \times 100\% \tag{4.3}$$

$$\text{Relative Bias of } Pr(T=0) = \left|\frac{Pr_{est}(T=0) - Pr_{true}(T=0)}{Pr_{true}(T=0)}\right| \times 100\% \tag{4.4}$$

Methods	Exp(T)		Var(T)		Pr(T)	
	Average	Relative Bias(%)	Average	Relative Bias(%)	Average	Relative Bias(%)
True Value	5.67		15.30		0.071	
Solomon	5.68	0.102	14.93	2.417	0.072	0.888
Bootstrap I	4.70	17.038	15.56	1.717	0.082	15.588
Bootstrap II	5.69	0.407	14.78	3.365	0.075	5.168
Bootstrap III	5.69	0.279	14.98	2.090	0.072	1.659

Table 4.1: Simulation Results from Normal Distribution

The simulated results and relative biases are listed in Table 4.1 – Table 4.4. We summary the results as below:

CASE 1, normal distribution: In this case, the model assumptions are met. We see that the relative biases of the estimates of $E(T)$ and $V(T)$ using Solomon's parametric method are very small, as expected (see table 4.1). Looking at the results from bootstrap methods, the third bootstrap method (bootstrapping on the transformed data, and then computing estimates from the inverse transformed bootstrap samples) also gives a very good result, almost as good as Solomon's parametric method. This shows that the third bootstrap method retains the same accuracy when the model assumptions are met.

Methods	Exp(T)		Var(T)		Pr(T)	
	Average	Relative Bias(%)	Average	Relative Bias(%)	Average	Relative Bias(%)
True Value	5.79		21.79		0.135	
Solomon	5.79	0.068	21.07	3.310	0.133	1.405
Bootstrap I	4.91	15.176	20.13	7.642	0.121	10.709
Bootstrap II	5.81	0.283	20.78	4.629	0.138	1.609
Bootstrap III	5.80	0.051	21.07	3.313	0.134	0.790

Table 4.2: Simulation Results from Mixture Normal Distribution

Methods	Exp(T)		Var(T)		Pr(T)	
	Average	Relative Bias(%)	Average	Relative Bias(%)	Average	Relative Bias(%)
True Value	0.52		0.76		0.668	
Solomon	1.22	134.740	1.10	44.707	0.334	50.058
Bootstrap I	0.79	51.554	0.76	0.119	0.579	13.316
Bootstrap II	1.16	123.877	1.06	39.254	0.361	46.037
Bootstrap III	0.73	39.827	0.95	25.823	0.589	11.791

Table 4.3: Simulation Results from Cauchy Distribution

Methods	Exp(T)		Var(T)		Pr(T)	
	Average	Relative Bias(%)	Average	Relative Bias(%)	Average	Relative Bias(%)
True Value	5.20		16.30		0.076	
Solomon	5.69	9.491	15.05	7.672	0.073	4.436
Bootstrap I	4.30	17.258	15.56	4.526	0.099	29.827
Bootstrap II	5.70	9.631	14.88	8.711	0.076	0.040
Bootstrap III	5.27	1.421	16.57	1.687	0.098	28.493

Table 4.4: Simulation Results from Double Exponential Distribution

CASE 2, mixture normal distribution: In this case, the distribution is very close to the normal distribution. The performances of the methods are very similar to Case 1, but it seems the third bootstrap method yields smaller relative bias compared to the parametric method (see table 4.2).

CASE 3, Cauchy distribution: All the methods produce large relative biases for estimating the three quantities, except the first bootstrap method for $V(T)$ (see table 4.3). The parametric method is the worst one, and the second bootstrap method is the second worst method, since both methods depend on the assumption of normal distribution. The third bootstrap procedure has better performance than other methods, according to the criteria of relative bias.

CASE 4, double exponential distribution: When estimating $E(T)$ and $V(T)$, the third bootstrap procedure has much smaller relative bias than all the other methods. The parametric method and the second bootstrap method have almost the same performance. But for estimating $P(T)$, the performance of these methods are somewhat different from estimating $E(T)$ and $V(T)$, the third and the first bootstrap method has a larger bias.

Now we present out final conclusions and remarks based on the findings from the above simulation study:

(1) Because both the model assumption and the analytical form of the formula of Solomon's method depend on normal distribution, it works best only under the normal distributional case. For other distributional cases, the relative biases are much larger than the third bootstrap method.

(2) The first bootstrap method, which uses external bootstrap weights, generally has a larger bias than other methods, in almost all cases and for estimating all the three quantities.

(3) The second bootstrap method, which bootstraps on the transformed data and then plugs the bootstrap estimates of the model parameters in Solomon's parametric formulae, behaves much the same as Solomon's parametric method. This may be because when computing the three quantities, it still relies on Solomon's parametric formulae, which are derived under the normal assumption.

(4) In general the third bootstrap method is the best among these methods. When the model assumptions are met, it gives estimates almost as good as the parametric method. But when the model assumptions are not met, it still produces estimates with very small relative bias. Thus the third bootstrap procedure is very robust against distributional assumptions, and should be preferred when the exact distributions are not known.

(5) The estimation of $P(T)$ using bootstrap methods deserves more research. From the simulation study, we found that its behaviours are somewhat different from $E(T)$ and $Var(T)$.

4.2 Application to the IPPPSH data

In order to demonstrate how to use the parametric method and the three bootstrap methods developed in the previous chapters, we give an example in the section by applying these methods to a real data set given in Solomon (1989).

The data set on blood pressure from the International Prospective Primary Prevention Study in Hypertension (IPPPSH) is from 25 hypertension males receiving treatment regimes containing a betablocker. Each patient was measured quarterly for a period of 4 years, and both diastolic and systolic blood pressures were measured twice each time, thus we have 4 25×16 data sets, two for diastolic blood pressures and two for systolic blood pressures. The data sets

From diastolic measurements with thresholds 95, 100, and 105

Methods	$h = 95$			$h = 100$			$h = 105$		
	Exp	Var	Pr	Exp	Var	Pr	Exp	Var	Pr
Observed	4.8	18.5	0.16	2.1	7.2	0.40	1.0	2.3	0.64
Solomon	5.4	15.0	0.08	2.7	8.4	0.26	1.2	3.2	0.52
Bootstrap I	4.4	17.9	0.12	2.0	8.0	0.32	0.9	3.2	0.53
Bootstrap II	5.7	15.3	0.07	3.0	9.0	0.24	1.3	3.7	0.49
Bootstrap III	5.5	17.3	0.09	2.6	8.3	0.26	1.2	2.9	0.47

Table 4.5: Estimated Results from the IPPPSH diastolic data, on log-scale

can be downloaded from http://www.maths.adelaide.edu.au/people/psolomon. For simplicity, we average the two diastolic measurements and the two systolic measurements taken at each time, and also we ignore the relationship between diastolic and systolic blood pressures, and analyze them separately. To reduce the influence of extreme values and make the assumption of normality more reliable, we do a log-transformation on the observed data. The observed values and fitted values using different methods and for a range of thresholds are listed in Table 4.5 and Table 4.6, where the observed values are computed directly from the data (see the formulae of (2.13), (2.14), and (2.15)). In tables 4.5 and 4.6, "Exp" denotes the expected value of number of exceedances, "Var" denotes the variance of the number of exceedances, and "Pr" denotes the probability of observing no exceedance.

From systolic measurements with thresholds 150, 160, and 170

Methods	$h = 150$			$h = 160$			$h = 170$		
	Exp	Var	Pr	Exp	Var	Pr	Exp	Var	Pr
Observed	6.2	26.8	0.12	4.0	21.8	0.24	2.2	13.8	0.52
Solomon	7.3	24.4	0.08	4.4	18.5	0.22	2.2	10.1	0.43
Bootstrap I	6.8	23.9	0.08	3.2	16.2	0.22	1.7	10.0	0.46
Bootstrap II	7.3	23.9	0.09	4.4	18.3	0.23	2.3	10.4	0.42
Bootstrap III	7.2	25.1	0.09	4.3	18.5	0.23	2.2	10.0	0.42

Table 4.6: Estimated Results from the IPPPSH systolic data, on log-scale

Looking at Table 4.5, we see that for expected number of exceedances, the estimated values are larger than the observed values, except the estimates using bootstrap method I. For the variance of number of exceedances, when the threshold is 95, all methods give a smaller estimate than observed, but when the thresholds are 100 and 105, their estimates are larger than the observed values. With the probability of observing no exceedance, the estimated values from all methods are smaller than the observed.

Examining the estimated results from the systolic data, we see that for the quantity of number of exceedances, when the threshold is 150 and 160, all the estimates are larger than observed value, except for bootstrap method I when the threshold is 160. When the threshold is 170, the estimates from Solomon's method, bootstrap method II and bootstrap method III are very close to the

observed value. Also we notice that in this case, Solomon's method, bootstrap method II and bootstrap method III produce very similar estimates. For the other two quantities, variance of number of exceedances and probability of observing no exceedances, all the methods give a smaller value than the observed value, and the estimated values are quite close.

Bibliography

[1] Bickel, Peter J. and Freedman, David A. (1981). Some asymptotic theory for the bootstrap. The Annals of Statistics, 9, 1196-1217.

[2] Efron,Bradley (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics. 7, 1-26.

[3] Efron,Bradley (2003). Second thoughts on the bootstrap. Statistical Science. 18, 135-140.

[4] Fuller. Wayne A. and Battese, George E. (1973). Transformation for estimation of linear models with nested-error. Journal of the American Statistical Association, 68, 626-632.

[5] Harris, Ian R. and Burch, Brent D. (2003). The probability of preponderancy, an alternative to the Intraclass Correlation. Technical Report, Department of Statistical Science, Southern Methodist University.

[6] Harris, Ian R. and Burch, Brent D. (2003). Measuring relative importance of sources of variation without using variance. Technical Report, Department of Statistical Science, Southern Methodist University.

[7] Henderson, C.R. (1953). Estimation of Variance and covariance components. Biometrics, 9, 226-252.

[8] Hu,Feifang and Zidek, James V. (1995). A bootstrap based on the estimating equations of the linear model. Biometrika, 82, 263-75.

[9] Lele,Subhash R (2003). Impact of Bootstrap on the Estimating Functions, Statistical Science, 18,185-190.

[10] Liu,Regina Y. (1988). Bootstrap procedures under some non-iid models. The Annals of Statistics, 16, 1696-1708.

[11] Rao, C.R. (1970). Estimation of heteroscedastic variances in linear models. Journal of American Statistical Association, 65, 161-172

[12] Rao, C.R. (1971).Estimation of variance and covariance components-MINQUE theory. Journal of Multivariate Analysis. 3, 257-275.

[13] Rao, C.R. (1972). Estimation of variance and covariance components in linear models. Journal of American Statistical Association, 67, 112-115.

[14] Searle,Shayle R., Casella,George, and McCulloch, Charles E. (1992), Variance Components, Wiley, New York.

[15] Shao,Jun and Tu,Dongsheng (1995). The Jackknife and Bootstrap, Springer Verlag, New York

[16] Singh, Kesar (1981). On the asymptotic accuracy of Efron's bootstrap, The Annals of Statistics, 9, 1187-1195.

[17] Solomon, P.J. (1989). On components of variance and modelling exceedances over a threshold. Austral. J. Statist., 31, 18-24.

[18] Thach, Thuan (1998). Some contributions to Bootstrap and empirical likelihood methods for some non-i.i.d models. PhD thesis, University of Alberta.

[19] Wu, C.F.J (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis, The Annals of Statistics, 14, 1261-1295.