# Flexible Modeling For Hierarchical Data, Data With Random Sample Sizes and Selection Bias, with Applications in Pharmaceutical Research

**Elasma Immaculate Milanzi**

Promotor: Prof. dr. Geert Molenberghs

Co-Promotor: Prof. dr. Ariel Alonso

## Jury List

Prof. dr. Geert Molenberghs (Promoter)
Universiteit Hasselt & Katholieke Universiteit Leuven, BE

Prof. dr. Ariel Alonso (co-promoter)
Maastricht University, NL

Prof. dr. Tomasz Burzykowski (adv. Committee member)
Universiteit Hasselt & IDDI, BE

Prof. dr. Michael G. Kenward
London School of Hygiene and Tropical Medicine, UK

Prof. dr. Geert Verbeke
Katholieke Universiteit Leuven & Universiteit Hasselt, BE

Prof. dr. Gerard van Breukelen
Maastricht University, NL

Dr. Luc Bijnens
Janssen Pharmaceutica & Universiteit Hasselt, BE

September 26, 2013

# Samenvatting

Ruw geschetst bestaat het ontwikkelingsproces van nieuwe geneesmiddelen uit de volgende stappen: de ontdekkingsfase, waar potentieel actieve chemische componenten worden onderscheiden die verdere studie vereisen; de optimisatie-fase die de farmacologische profielen optimiseert, en de ontwikkelingsfase waar de potentiële component aan rigoureuse evaluatie wordt onderworpen. Het is uiteraard belangrijk dat het finale product veilig en werkzaam is, binnen de populatie die men voor ogen heeft (Schultz, Ruppel, and Johnson, 1988). Gemeenschappelijk aan alle fasen is het gebruik van empirische evidentie, of gegevens, om het proces en de eraan gekoppelde beslissing te ondersteunen. Er is dus grote nood aan statistische expertise. De klemtoon hier ligt op het ontwikkelen van gepaste methodologie, gekoppeld aan ingewikkeld proefopzet, in de ontdekkings- en ontwikelingsfases. Ze vormen het onderwerp van respectievelijk Deel I en Deel II van onderhavig werk.

## Flexibele methodologie voor hiërarchische gegevens, en voor gegevens met selectie-effecten

Farmaceutische bedrijven houden bibliotheken bij van voor de ontwikkeling van geneesmiddelen veelbelovende chemische componenten. Het is cruciaal dat dergelijke bibliotheken een grote fractie "interesssante" componenten bevatten. Dit verhoogt uiteraard de kans op succes bij screening (Lajiness and Watson, 2008). Het is gebruikelijk van de eigen ontdekkingen aan te vullen met aangekochte bibliotheken. Recent werd voorgesteld van de bibliotheken te versterken door ze te voorzien van de opinie van experten (Hack *et al.*, 2011).

De aanpak voorgesteld door Hack *et al.* (2011) vertrekt van de aankoop van verscheidene structurele filters die ook de eigenschappen van de componenten screenen. Hierdoor is het mogelijk van onmiddellijk die componenten te verwijderen die geen

enkele belofte vertonen. De resterende componenten worden dan in zogenaamde clusters ondergebracht, samen met de reeds in huis aanwezige componenten. Clusters die uitsluitend uit externe componenten bestaan worden voorgelegd aan de wereldwijde gemeenschap van medicinale chemici; zij scoren de componenten om op die manier uit te maken of ze een plaats verdienen in de bibliotheken of niet. Naast een ja/neen beslissing worden de componenten ook van een rangorde voorzien, uiteraard met het oog op het aanbrengen van prioriteiten.

Het boven geschetste proces heeft hoog-dimensionale aspecten om twee redenen: (i) als experten vele clusters scoren, dan is de dimensie van de respons vector hoog; (ii) een score toekennen aan een cluster impliceert het schatten van duizenden fixed-effect parameters.

Uiteraard is de methodologie voor hiërarchische gegevens goed ontwikkeld (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000; Liang and Zeger, 1986). Er is heel wat vooruitgang geboekt ook bij de analyse van hoog-dimensionale herhaalde metingen. Bijvoorbeeld, Fieuws and Verbeke (2006) maken gebruik van paarsgewijs schatten, terwijl Molenberghs, Verbeke, and Iddi (2011) grote steekproeven in stukjes hakken, elk stukje apart analyseren, en dan volgens bepaalde combinatieregels tot één conclusie komen.

Om expert opinie te kwantificeren is het nodig van de bestaande methodologie uit te breiden zodat tegelijkertijd de beide hoog-dimensionale aspecten (fixed effecten en herhaalde respons vector) in rekening kunnen gebracht worden. Een dergelijke procedure wordt voorgesteld in Hoofdstuk 3. Vertrekkend van de splitsingsidee in Molenberghs, Verbeke, and Iddi (2011), wordt een *permutatie-splitsing procedure* voorgesteld. Ze laat toe van het geschetste probleem aan te pakken binnen de grenzen van standaard beschikbare statistische software. De resultaten liggen zeer dicht bij de maximum likelihood schatters die men zou krijgen indien de steekproef als geheel wordt geanalyseerd. Alleen is er een enorme winst aan berekeningstijd en -vereisten. Dit is mogelijk door: (i) oordeelkundig splitsen van de dataset is deelverzamelingen; (ii) adequate schattingsmethoden toepassen op elk van de delen; (iii) permutatie van de gegevens en herhalen van stappen (i) en (ii); (iv) combinatie van de voor de delen verkregen schatters tot één enkele conclusie. De performantie van de methode wordt ook onderzocht aan de hand van simulaties.

In deze methode is het niet zo dat het aantal clusters dat door een expert behandeld wordt bij voorbaat vast ligt. In overeenstemming met de praktijk hangt zulks af van de tijd beschikbaar voor een bepaalde expert. Het aantal bestudeerde clusters (*number of clusters rated, NoCR* bevat meer dan waarschijnlijk minstens een beetje informatie over de scores van de expert. In Hoofdstuk 4 worden de theoretische im-

plicaties hiervan besproken. Het belang van het mee in rekening brengen van NoCR wordt aangetoond, zelfs onafhankelijk van het feit of het al of niet een invloed heeft op de score van een expert. Daarnaast worden aantrekkelijke proefopzetten besproken die dit probleem vermijden, zoals dat waarbij een expert alle clusters bestudeerd, of het random toekennen van een aantal clusters over de experten, waarbij het aantal wel degelijk wordt vastgehouden. Ondanks hun theoretische voordelen zijn ze voor de praktijk minder aangewezen. Pragmatisch kan het dus niet anders dan toch maar met NoCR rekening te houden.

De meeste methoden voor niet-gerandomiseerde studies impliceren een vorm van data-verrijking (*enrichtment*). Dit betekent dat er meer in het model verondersteld wordt dan gegevens kunnen valideren. Verrijking stoelt dus op niet-verifieerbare aannames. Typische voorbeelden van verrijking zijn: ontbrekende gegevens, censurering bij overlevingstijden, random effecten, enz. Het foutief specifiëren van de random effect verdeling kan problemen veroorzaken voor de statistische conclusies (Litière, Alonso, and Molenberghs, 2008). Om die reden zoeken we naar methodologie die robuust is tegen misspecificatie, omdat verrijking nu eenmaal niet te vermijden is. Hoofdstuk 5 stelt een dergelijke methode voor. De impact op de conclusies wanneer dit fenomeen verwaarloosd wordt, vormt het voorwerp van studie in Hoofdstuk 6. Via simulaties wordt ook nagegaan wat er gebeurt indien overdispersie wordt verwaarloosd.

## Flexibele methodologie voor gegevens met random steekproefgrootte

Klinische studies gaan na of een potentieel geneesmiddel voldoende veilig en werkzaam is (Rodda *et al.*, 1988). Om de impact op de studiepopulatie te verkleinen, maakt men sedert decennia gebruik van zogenaamde random steekproefgrootte (*random sample size, RSS*). Dit heeft geleid tot het kader van de groep sequentiële studies (*group sequential trials, GST*). Een GST kan gestopt worden indien het resultaat vroeg in de studie buiten verwachting heel sterk zou zijn, of wanneer net het tegendeel voorkomt. Er zijn duidelijk ethische en economische voordelen aan deze manier van werken, maar tegelijk zijn er problemen op het vlak van parameterschatting. Er is een brede consensus dat schatters gebaseerd op GST minder elegante eigenschappen hebben dan wanneer conventionele gegevens uit een studie met vaste steekproefgrootte gebruikt worden. Bijvooorbeeld, het steekproefgemiddelde (*sample average, SA*) verliest de zogenaamde minimum variantie onvertekende eigenschap (Todd, Whitehead, and Facey,

1996; Jennison and Turnbull, 2000). Als antwoord hierop werden heel wat alternatieve schatters voorgesteld (Whitehead, 1997; Emerson and Fleming, 1990; Liu and Hall, 1999). Deel Part II bestudeert dit probleem in detail en vanuit een orginele invalshoek. Ten eerste wordt RSS gekoppeld aan het nu goed ontwikkelde gebied van *joint modeling*, waarbij ook de link gelegd wordt met onvolledige gegevens en overlevingsanalyse. Concepten zoals *ignorability*, *separability*, en *ancillarity* kunnen dan handig binnen deze context geplaatst worden om op die manier eigenschappen van lineaire schatters af te leiden. De relevantie voor het kader van de klinische studie wordt bestudeerd door de nadruk te leggen op data uit GST. We leiden af dat standaardschatters een veel grotere geldigheid binnen de context van GST dan meestal wordt aangenomen. Een en ander wordt grondig bestudeerd in Hoofdstukken 7 en 8 . Naast de hoger genoemde eigenschappen wordt ook het verband gelegd met statistische volledigheid, sufficiënte statistieken en de stelling van Lehman-Scheffé. Tenslotte is er een verband met concepten uit de ontbrekende gegevens, zoals *missing at random (MAR)* en *ingorability*. Een cruciaal gegeven is dat bijvoorbeeld het gewone steekproefgemiddelde nog steeds volgt uit het gebruik van maximum likelihood, ondanks het verlies van een aantal schijnbaar belangrijke frequentistische eigenschappen. Daarnaast wordt ook conditionele maximum likelihood gebruikt om een schatter af te leiden die onvertekend is ook in kleine steekproeven. Het verschil tussen de schatters verkregen uit de gewone en de conditionele likelihood is nauw verwant aan de vertekening die vanuit frequentistisch oogpunt bestudeerd wordt.

De meeste geneesmiddelen die bedoeld zijn om het leven te verlengen worden ook bestudeerd in functie van kwaliteit van het leven. Dit laatste wordt vaak in kaart gebracht door het gebruik van gevalideerde schalen. Ze moeten natuurlijk geldig en betrouwbaar zijn, in de psychometrische betekenis van het woord; dit betekent dat ze voldoende precies datgene meten wat ze verondersteld worden van te meten. Indien de schalen een continue maat opleveren, is betrouwbaarheid uit te drukken als een verhouding van varianties. Voor binaire respons is dit minder evident. In Hoofdstuk 9 worden benaderende uitdrukkingen afgeleid voor de betrouwbaarheid in voorkomend geval; een en ander wordt binnen het *Item Response Theory* paradigma geplaatst. Gebaseerd op de benaderende, zogenaamd manifestie correlatiefuncties van Vangeneugden *et al.* (2010) kunnen we aantonen dat betrouwbaarheid van een binaire schaal evengoed als een variantieratio kan berekend worden. Dit vermijdt uiteraard belangrijke computationele problemen.

# Acknowledgments

It is an open secret that this work is a product of immense collaborations from different angles.

Geert Molenberghs and Ariel Alonso: I cannot ask for a better team to work with. The knowledge and experience, both academic and non-academic that you have imparted to me is immeasurable and I will forever be grateful. Your sensitivity towards my responsibilities as a parent and making our working schedules as flexible as possible has largely contributed to the my success of the work being celebrated today.

My appreciation extends to Luc Bijnens, Christophe Buyck and colleagues from Janssen pharmacuetica with whom we worked together on the projects of the first part of the thesis.

A special recognition goes to the jury members for taking time to read the thesis and for the enlightening comments. Many thanks to my office mates in C107 and E104 and I-BioStat colleagues for offering the stimulating and habitable environment.
The financial support from BOF cannot be taken for granted.

My family in Malawi: Regardless of the distance, I feel your support close by, of course with Edith around its even much closer :-) . Thanks for always being there for me.

To my son Dalitso, I will make up for all the missed scout/swimming sessions and thanks for frustrating me with other things different from non-converging simulations due to a forgotten comma :-). Having you around is just the best.

My belgian family(Sonia, Valere, Amanda, and Karin): Bedankt voor alles. Jij bent zoals familie. Many thanks to the kenyan students community for your friendship.

Njeru Njagi: I cannot trade your companionship for anything, it makes a lot of things lighter and bearable. Asante sana.

My special tribute to Arthur Gitome. We were looking forward to this day together. May his soul rest in peace.

# List of Publications

This work has been based on the following scientific papers:

Alonso, A., **Milanzi, E**., Buyck, C., Molenberghs, G. and Bijnens, L. (2013). Impact of selection bias on the qualitative assessment of clusters of chemical compounds. Submitted.

Alonso, A., **Milanzi, E**., Buyck, C., Molenberghs, G. and Bijnens, L. (2013). A new modeling approach for quantifying expert opinion in the drug discovery process. *Submitted, revised.*

**Milanzi, E**., Alonso, A.,Buyck, C., Molenberghs, G. and Bijnens, L. (2013). A permutational-splitting sample procedure to quantify expert opinion on clusters of chemical compounds using high-dimensional data. *Submitted, revised.*

**Milanzi, E**., Alonso, A. and Molenberghs, G. (2012). Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. Statistics In Medicine 2012; 31, 14751482.

**Milanzi, E**., Molenberghs, G., Alonso, A., Kenward, M.G., Verbeke,G.,Tsiatis, A. A. and Davidian, M. (2013). Properties of Estimators in Exponential Family Settings With Observation- based Stopping Rules. *Submitted, revised.*

**Milanzi, E**., Molenberghs, G., Alonso, A., Kenward, M. G., Verbeke,G.,Tsiatis, A. A. and Davidian, M. (2013). Estimation after group sequential trials. *Submitted, revised.*

**Milanzi, E**., Molenberghs, G., Alonso, A., Verbeke,G., and De Boeck, P. (2013). Reliability Measures In Item Response Theory: Manifest vs Latent Correlation Functions. *Submitted, revised.*

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| 1PL | One Parameter Logistic |
| 2PL | Two Parameter Logistic |
| BAM | Bias Adjusted Estimator |
| CI | Confidence Interval |
| CL | Conditional Likelihood |
| CLE | Conditional Likelihood Estimate |
| CRSS | Completely Random Sample Size |
| CTT | Classical Test Theory |
| FSS | Fixed Sample Size |
| GLMM | Generalized Linear Mixed Model |
| GSA | Generalized Sample Average |
| GST | Group Sequential Trials |
| GT | Generalizability Theory |
| ICC | Intra-Class Correlation |
| IRT | Item Response Theory |
| LSAT | Law School Admission Test |
| MAR | Missing At Random |
| MCAR | Missing Completely At Random |
| MLE | Maximum Likelihood Estimate |
| MNAR | Missing Not At Random |
| MSE | Mean Square Error |
| MUE | Mean Unbiased Estimator |
| NoCR | Number of Clucters Rated |
| RBADJ | Rao's Bias Adjusted |
| RSS | Random Sample Size |
| SA | Sample Average |

# Chapter 1

# Introduction

Introduction Roughly, the drug discovery process is divided into three stages, namely, lead discovery, lead optimization, and lead development, in that order. The lead discovery stage identifies potentially active chemical compounds worth of further study for drug development, and the lead optimization stage improves the pharmacological profiles of the identified compounds, by increasing the level of desirable activity and reducing the level of undesirable activity. Finally, the lead development stage, subjects the compounds to rigorous evaluations, to ensure that the end product is both safe and effective for the targeted population (Schultz, Ruppel, and Johnson, 1988). Common to all stages and critical part of evidence qualifying compounds to proceed to the next stage, is the amount of empirical evidence in support of a decision. This makes statistical expertise indispensable in the whole drug discovery process. The thesis focuses on flexible methodology and complications encountered during the analysis of empirical data realized during compound the acquisition process in lead optimization stage and clinical trials in the lead development stage. These are the focus of Parts I and II, respectively.

## 1.1 Flexible Methodology for Hierarchical Data and Data with selection Bias

Pharmaceutical companies tend to maintain a library of chemical compounds that are known to possess drug-like matter. Regularly, these are screened for drug-activity, to identify compounds that can be developed further. It is crucial that such a library contain a large proportion of "interesting" compounds, from a pharmaceutical

point of view, to increase chances of hits during screening (Lajiness and Watson, 2008), and this is usually achieved by supplementing the library collection through acquisition of compounds from vendors. Prevalent techniques used in determining compounds worth acquiring,are frequently based on similarities or differences in properties between compounds already in the library and those to be acquired (Ma, Lazo, and Xie, 2011; Dunbar, 2000). However, Hack *et al.* (2011) recently suggested an approach for enhancing diversity of a chemical library that incorporates expert opinion as additional evidence when deciding on which compounds to acquire. Statistical challenges and possible solutions associated with quantification of the expert opinion, which include a combination of data hierarchy and high-dimensionality, selection bias and/or missing data, are extensively addressed in Part I of this thesis.

The approach suggested by Hack *et al.* (2011) proceeds by screening candidate compounds for acquisition using various structural and property filters in order to eliminate clearly non-drug-like matter. The remaining compounds are then clustered together with the in-house collection using a novel fingerprint-based clustering algorithm that emphasizes common substructures and works with millions of molecules. Clusters populated exclusively by external compounds are identified as "diversity holes", and the representative members are presented to the global medicinal chemistry community, to rate the clusters as to whether they should be included in the library or not. Finally, the ratings are quantified and used to rank the clusters according to acquisition priority.

Though the approach seems straightforward, its implementation poses statistical modeling challenges. Since each expert can rate more than one cluster, the collected ratings have a hierarchical structure, hence standard methods that assume independence typically do not apply. Chemical compounds are usually acquired in millions, although in the approach considered here, this was reduced to thousands (22,015), through clustering. Certainly, high-dimensional problems crop up from two angles, namely, (i) when an expert rates many clusters, the dimension of the repeated response vector will be high, and (ii) when assigning a rank to each cluster implies the estimation of thousands of fixed effects parameters.

Indeed, methodology for hierarchical data is well established (Molenberghs and Verbeke, 2005; Verbeke and Molenberghs, 2000; Liang and Zeger, 1986) , and there are remarkable research advances in high-dimensional data problems. To circumvent the problem of high-dimension repeated response, Fieuws and Verbeke (2006) employ a pairwise fitting procedure, while Molenberghs, Verbeke, and Iddi (2011) use sample

splitting. While relevant, these methods do not fully tackle the problem of a high-dimensional fixed effects vector.
In regression models, variable selection Meinshausen and Bühlmann (2010); Fan and Peng (2004) precedes model fitting to avoid a high-dimensional fixed-effects vector. Unfortunately, variable selection is not a viable option since each cluster has to be ranked. Evidently, to quantify expert opinion existing methods need to be extended to simultaneously address high-dimensionality in both the fixed-effects and repeated response vectors, and Chapter 3 proposes such a procedure. Building on the sample splitting idea, the *permutational-splitting sample* procedure, allows for model estimation for data with the two-fold high-dimension problem, using the commonly available computing resources and produces estimates similar to maximum likelihood estimates that would be obtained if the whole dataset was used at once. This is achieved by, (i) judiciously splitting the whole dataset into sub-samples, (ii) performing valid model estimation for each sub-sample, (iii) permuting the data and repeating (i) and (ii), and (iv) mechanically combining the parameter estimates from the sub-sample analyses to obtain the final estimates. In addition to quantifying expert opinion using this procedure, Chapter 3 presents a simulation study that investigates the performance of the procedure against full maximum likelihood.

In this approach, the clusters were presented in a random order, but the final number of clusters each expert rated was not fixed in advance, rather it depended on the time an expert decided to stop rating. Indeed, the number of clusters rated (NoCR), contains some information about the expert and thus may have to be taken into account when quantifying expert opinion. Chapter 4 discusses the theoretical implications of having differing and random NoCR, and demonstrates why it is important to consider the process generating the NoCR, regardless of whether it influences the rating outcome or not. Further, appealing designs that avoid such a phenomenon, namely, each expert rating all available clusters or randomly assigning a pre-fixed number of cluster and force the experts to finish their quota, are explored. In spite of their statistical appeal, their practicality is doubtful. As such, to quantify the expert opinion, methods that can model the rating and NoRC generating processes jointly, unlike methods that assume randomization, have to be used.

Most methods for non-randomized studies require data enrichment, which is mainly based on unverifiable assumptions. Common forms of enrichment include random effects and (non)parametric models, which if misspecified can have detrimental effects on estimates and statistical inferences (Litière, Alonso, and Molenberghs, 2008). It is desirable that methods that are robust against misspecification are used,

especially in situations where data enrichment is unavoidable. Chapter 5 therefore proposes a modeling technique that uses random effects to augment the data and produces valid estimates when the assumptions are misspecified. Using the technique introduced in Chapter 5, Chapter 6 studies the impact of ignoring data augmentation when it is necessary. Through an extensive simulation study, problems resulting from ignoring overdispersion, an example of a situation where data enrichment is necessary, are exposed and possible solutions suggested.

## 1.2 Flexible Methodology For Data With Random Sample Size

Clinical trials fall into the lead development stage, where the value of a compound as drug, in prevention, treatment or diagnosis of a disease is determined by evaluating its benefits relative to its risk and undesirable effects (Rodda *et al.*, 1988). In view of minimizing the impact of undesirable effects on the tested population, study designs with random sample size (RSS), like group sequential trials (GST), which can be stopped in case of clear danger or benefit are preferred. While convenient ethically, such designs pose statistical challenges in estimating parameters of interest. The consensus is that estimators used after a fixed sample size (FSS) trial, lose some of their nice properties when used after a GST. For example, the sample average (SA), loses the minimum variance unbiased estimator property (Todd, Whitehead, and Facey, 1996; Jennison and Turnbull, 2000). This has led to research directed at finding alternative estimators to be used after GST (Whitehead, 1997; Emerson and Fleming, 1990; Liu and Hall, 1999). Part II presents an in-depth analysis of properties of estimators for studies with (RSS) from a novel perspective. First, RSS is linked to the well-established research area of joint modeling, which includes settings like incomplete data and survival analysis. Then, concepts like ignorability, separability and ancillarity are placed in the context of RSS trials theory to establish properties of linear mean based estimators. Further the relevance of such properties to clinical trials is illustrated by studying the specific case of GST where it is shown that, while retaining the good qualities of RSS, flexible analyses like those used in FSS trials may be adequate after a GST.

When the human population is the target, clinical trials are more likely to expose the participants to some risk or unexpected liability. To this effect, strict ethical guidelines in medical research like, *"The Helsink 1964"* and *"The Numbero code"* are enforced to minimize such risks. For example, it is unacceptable to continue

subjecting participants in a study to a clearly toxic or less effective drug when a better alternative drug exists (Armitage, 1975). This forms the basis of most clinical trials that are designed to allow early stopping. In addition to adhering to high-level ethics, such trials are economical in resources. Recall that these attractive features come at a cost of losing some good properties of mean estimators in the FSS setting, hence casting doubt on their use in the RSS setting. By relating RSS trials to missing data theory, Chapter 7 clearly outlines important concepts, common to all these settings, that determine the properties of mean estimators in RSS trials. In particular, completeness of a sufficient statistic and Lehman-Scheffe's theorem are used to illustrate the loss of unbiased minimum variance mean estimator (UMVUE) property by SA. Further, concepts like the missing at random (MAR) and ignorability assumptions from the missing data theory are used to show that sample average (SA) is the maximum likelihood estimator (MLE) after a RSS trial. As such likelihood inferences valid under MAR in the missing data context, like asymptotic unbiasedness of MLE and validity of asymptotic confidence intervals computed using observed information matrix, are also valid in RSS trials. SA is also studied within the larger class of linear based mean estimators where it emerges as the only estimator of that class that exhibits asymptotic unbiasedness. An interesting result is that, mean estimators used in FSS trials are still useful in RSS trials. Finally, a new unbiased conditional maximum likelihood mean estimator is introduced as an alternative for those not comfortable with the finite sample bias in MLE, although its use is accompanied with some loss of information.

Chapter 8 digests the relevance of the findings in Chapter 7 to GST, which is a specific case and a popular design of RSS trials. The main message is that flexible modeling used in FSS trials is useful in GST, when the likelihood estimation paradigm is followed. For example, SA can still be used as a mean estimator despite having finite bias.

Evaluation of most drugs meant to prolong life, include quality of life assessment studies, where quality of life is measured according to an approved scale. Of the many aspects considered when approving the scale, is reliability, defined as the ability of the scale to consistently measure what it is supposed to measure. For scales with continuous scores, reliability is calculated as the ratio of true to observed variance, and is easily obtained. However, for binary scores, computing reliability as the ratio of true to observed variance is computationally challenging. Chapter 9 introduces approximate reliability measures for binary scored scales, within the Item Response

Theory paradigm. Based on approximate manifest correlation functions proposed by Vangeneugden *et al.* (2010), we show that reliability of a binary scored scale can be obtained as the ratio of true to observed variance, while eluding the computational obstacles.

# Chapter 2

# Motivating Case Studies

In this chapter, we introduce case studies that motivated the ideas developed in Part I and used to illustrate the findings in part II.

## 2.1 Expert Opinion On Clusters of Chemical Compounds

The pharmaceutical company Johnson & Johnson carried out a project to identify compounds from vendors for acquisition, to enhance the diversity of their library collection. The $5,261,676$ compounds marked for potential acquisition, underwent a filtering process to eliminate clearly non-drug-like matter. Remaining compounds were combined with the existing library collection, and a fingerprint-based procedure was implemented to cluster together related compounds. In total, $22,015$ clusters were exclusively made up of compounds from vendors, and viewed as having library diversification potential. To decide on which clusters to give acquisition priority, global medicinal experts were asked to rate these clusters.

### 2.1.1 Design of the study

The rating system was implemented by a desk-top application, Third Dimension Explorer (3DX), a drug-discovery specific software that is similar in concept to Microsoft Excel (Agrafiotis *et al.*, 2007). In a typical rating session, an expert was presented with a random small subset of clusters, selected from the entire set of $22,015$ clusters, to ensure that an expert looks at all clusters without scrolling through the screen. To help the expert make an informed choice, the clusters were presented with additional

information that included the size of the cluster, the structures of its representative members, i.e., the compound with the lowest molecular weight, and up to four additional members (the compound with the highest molecular weight and 1–3 other randomly chosen members). Possible ratings were: $-1$, if disliked, 0, if indifferent and 1 if liked, but for the analysis $-1$ and 0 were merged and recoded into 0. A new random subset that excludes clusters already rated, would be presented either, when all the clusters in the previous subset have been rated, or if the expert resumes the rating after a break. Rates were assigned to clusters and not individual members, and a total of 147 experts took part in the study.

The histogram in Figure 2.1 displays the distribution of the number of clusters rated by the experts. The left side distribution, which plots numbers of all experts, is positively skewed, indicating that many experts opted to rate few clusters. Indeed, 25% of the experts rated fewer than 345 clusters, 50% fewer than 1200 and 75% of the experts, fewer than 2370 clusters. Moreover, the most rated cluster had 31 ratings, and 8 for the least rated. On the right hand side is the distribution of the number of clusters for experts who rated less than 4000 clusters and it has two notable peaks at 0–200 and 2000 clusters, suggesting that many experts' number of ratings fall into these categories. In total, the final dataset contained 409,552 observations.

### 2.1.2 Possible Implications Of The Design On Estimation and Inference

The rating system was designed to support multiple sessions that would allow the experts to stop and resume the rating at their own convenience, resulting in numbers of clusters rated by experts ranging from 20 to 22,015. While practically convenient, it may bring about serious complications for the data analysis, especially when estimating the success probabilities. Assuming that each expert was expected to rate all clusters, vectors of ratings for those who did not achieve this can be considered incomplete, and missing data techniques can be used to account for the differing numbers. The magnitude of the missing responses (87 %), may require a large scale sensitivity analysis. Alternatively, the ability to stop may encourage experts to stop when they encounter a hard-to rate cluster so as to get a new random subset, in which case the wide range for the number of clusters rated would reflect selection patterns of the experts that translate into selection bias in estimation. In such a setting, selection bias methods can also be employed.

Serious complications can also arise from the less restricted random assignment of the clusters. Possible extreme cases include: (i) some clusters being rated by all

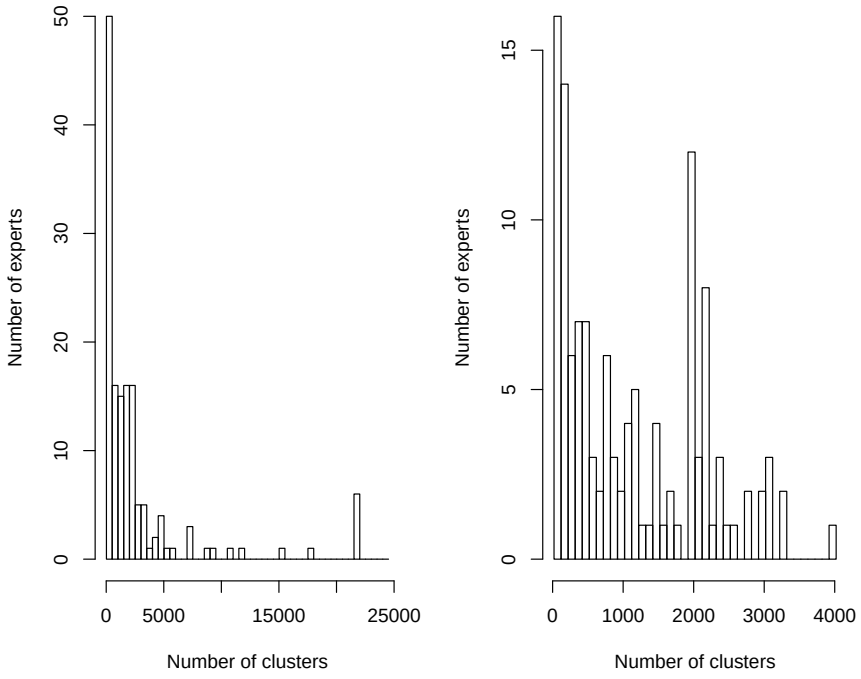


Figure 2.1: *Histogram for the number of clusters rated by the experts.The left hand side, is the histogram for all experts, and the rights hand side is for experts who rated less than 4000 clusters.*

experts and some by none, making assignment of ranks impossible, (ii) both experts and clusters being divided into two separate components, where one half of experts rates one half of the clusters, which may complicate estimation of experts' variability. Fortunately, the extreme cases were avoided since four experts rated all clusters. Nevertheless, the most rated cluster had $\approx 4\times$ the ratings of the least rated, i.e., the success probability for the former is estimated more precisely since it has more information.

The ideal design is to force the experts to rate all clusters, which may be impractical due to time constraints and conflicting engagements. Alternatively, each expert would be randomly assigned a subset of clusters in a way that ensures that each cluster is rated equally, and this requires that each expert finishes his/her quota of clusters which is also challenging. Even though it is not the focus of the present work, it is clear that the design of the study is another important element to guarantee the validity of the results. Optimal designs are a class of experimental designs that are optimal with respect to some statistical criterion (Berger and Wong, 2009). For instance, one may aim to select the number of experts, the number of clusters assigned to the experts and the assignment mechanism to maximize precision when estimating the probabilities of success. In principle, it seems intuitively desirable for each cluster to be evaluated by the same number of experts and for each pair of experts to have a reasonable number of clusters in common. However, more research will be needed to clarify these issues and establish the best possible design for this type of studies.

## 2.2 Epilepsy Data

The data come from a randomized, double-blind, parallel group and multi-center clinical trial for the comparison of placebo with a new anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in Faught *et al.* (1996). Randomization took place after a 12-week baseline period that served as a stabilization time for the use of AED's, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group and 44 to the active (new) treatment group. Patients were then measured weekly and after a followed up of 16 weeks (double-blind) they were entered into a long-term open-extension study. Consequently, some patients were followed for up to 27 weeks. The outcome of interest was the number of epileptic seizures experienced during the most recent week. The research question was whether or not the new treatment could reduce the number of epileptic seizures.

## 2.3 Verbal Agression Data

The data consist of subjects' responses to questions about verbal aggression. The instrument is a behavioral questionnaire. All items refer to verbally aggressive reactions in a frustrating situation. The data can also be considered as from a psychological experiment which has three design factors: (1) Behavior mode: a differentiation is made between two levels, i.e., wanting to do and the actual doing; (2) Situation type: This factor has two levels, namely other-to-blame and self-to-blame situation type, and each of these levels has two situations. Self to blame situations were: *'The grocery closes just as I am about to enter'* and *'The operator disconnects me when I had used up my last 10 cents'*. Other to-blame situations were: *'A bus fails to stop for me'* and *'I miss the train because a clerk gave me faulty information'*. So, the situations can also be viewed as nested in the situation type; (3) Behavior type: this had three kinds of behaviors, namely shout, scold, and curse. An example of an item in this instrument was: *'A bus fails to stop for me. I would want to curse'*. Possible answers were, no (0), perhaps (1), and yes (2). In our application, we will use the dichotomized version of the response, in which 'no' and 'perhaps' are recoded as 0 and 'yes' as 1. A detailed description of the data and its items can be found in Vansteelandt (2000) and De Boeck and Wilson (2004).

## 2.4 Law School Admission Test (LSAT6) Data

LSAT is a standardized test administered to prospective law student and designed to assess reading comprehension, logical and verbal reasoning proficiencies. The data comprises scores on five items of Section 6 of of the LSAT for 1000 examinees. The data is publicly available in the R package *mirt*, and it is well described in Bock and Lieberman (1970).

# Part I

# Flexible Methodology For Hierarchical Data and Data with Selection Bias

# Chapter 3

# A Permutational-Splitting Sample Procedure to Quantify Expert Opinion on Clusters of Chemical Compounds Using High-Dimensional Data

## 3.1 Introduction

The lengthy and expensive process of drug development is initiated with the lead discovery stage, which identifies potentially active chemical compounds worth of further study for drug development. Pharmaceutical companies tend to maintain a library of chemical compounds (library) that are screened for some drug activity. Lajiness and Watson (2008) advocate for a library with a large proportion of "interesting" chemical compounds, from a pharmaceutical point of view, to increase chances of hits during screening. Acquisition of third party chemical compounds ('compounds') presents a possibility to build such a library, though it comes with the challenge of selecting the compounds worth purchasing. Critical in selecting such compounds is the amount of

evidence supporting the presence of drug activity.

Recently, Hack *et al.* (2011) introduced an approach for enhancing the diversity of a library based on the theory of wisdom of crowds (Surowieck, 2004), when acquiring compounds from vendors. First, candidate compounds for acquisition are screened using various structural and property filters to eliminate clearly non-drug-like matter, then the remaining compounds are clustered together with the compounds already in the library, using a fingerprint-based clustering algorithm. Finally, clusters populated exclusively by third party compounds are identified and presented to the global medicinal chemistry experts, who rate the clusters regarding their appropriateness for library inclusion. Based on the ratings, each cluster is ranked and the top ranked ones are given acquisition priority. Using expert opinion has been acknowledged as crucial element for judgment (Oxman, Lavis, and Fretheim, 2007).

This chapter shows that, based on these qualitative ratings and using hierarchical models, a probability of *success* (recommending a cluster for inclusion) can be assigned to each cluster. The main issue in this process is that the presence of several judges and many clusters lead to a high-dimensional vector of repeated responses and a high-dimensional fixed-effect structure as well.

Facets of the so-called *curse of dimensionality* (Donoho, 2000), in statistical estimation and inference are numerous, and constitute a substantial proportion of active statistical research. For instance, in multiple linear regression, Gaure (2013) and Guimaraes and Portugal (2010) studied this problem when a large number of covariates are included in the model. Likewise, Fieuws and Verbeke (2006) have proposed approaches to fit multivariate hierarchical models in settings where the responses are high-dimensional vectors of repeated observations.

Arguably, variable selection is the most recognized form of high-dimensional data problems (Fan and Peng, 2004; Meinshausen and Bühlmann, 2010; Fan, Guo, and Hao, 2012), where the number of explanatory variables is much larger than the sample size. The challenge is to select useful variables from a multitude of mostly "noisy" variables. As such, many variable selection methods are based on the assumption that the high-dimensional vector of explanatory variables is sparse, and the methods are meant to identify those with the highest probability of having a non-zero effect. This approach is not plausible for our problem because in essence we only have one variable with numerous categories (resulting into a high dimensional fixed effects vector), such that even when the effect for some categories is zero we cannot omit them from the analysis.

The approach followed here is based on permuting and splitting the original data set into mutually exclusive subsets that are analyzed separately and the posterior

combination of the results from sub-analyses. It is aimed at rendering the use of random-effects models possible when there is a huge number of clusters and/or a large number of experts. In this setting, conventional maximum likelihood is not computationally feasible, and alternative strategies are needed.

Data splitting methods are not new in tackling high dimensional problems: Chen and Xie (2012) use a split-and-conquer approach to analyze extraordinarily large data in penalized regression. Fan, Guo, and Hao (2012) utilize data-splitting technique to estimate variance in ultrahigh dimensional regression. Molenberghs, Verbeke, and Iddi (2011), formulated a splitting approach when either the repeated response vector was high-dimensional or the sample size too large.

The scenario studied here, however, is radically different: both the repeated response vector and the vector of covariates are high-dimensional. This requires a different splitting strategy, in which the covariates involved in each sub-sample are not the same and so are the estimated effects and Hessian matrices from each sub-analysis. Hence,the methods used by the above mentioned authors in combining estimates do not directly apply.

## 3.2 Estimating the Probability of Success

To facilitate the decision making process, it is desirable to summarize the large number of qualitative assessments given by the experts into a single probability of success for every cluster. One way to approach this problem is to use generalized linear mixed models. Alternatively, a simpler method is to use the observed probabilities of success, estimated as the proportion of ones that each cluster received. There are, however, good reasons to prefer the model-based approach. Indeed, hierarchical models bring more flexibility by allowing the inclusion of covariates associated with the clusters and the experts. They also permit extensions to incorporate the presence of selection bias or missing data and explicitly account for the fact that an expert may evaluate several clusters. In addition, the model-based approach naturally delivers an estimate of the inter-expert variability. Although it is not the focus of the analysis, a measure of heterogeneity among experts is a valuable element for the interpretation of the results and for the design of future evaluation studies.

To estimate the probability of success for every cluster, let us now denote the vector of ratings associated with expert $i$ by $\boldsymbol{Y}_i = (Y_{ij})_{j\in\Lambda_i}$, where $\Lambda_i$ is the subset of all clusters evaluated by the $i$th expert and $i = 1, \ldots, n$. A natural choice to model

these data is the logistic-normal model:

$$\text{logit}\left[P\left(Y_{ij}=1|b_i\right)\right]=\beta_j+b_i, \tag{3.1}$$

where $\beta_j$ is a fixed parameter characterizing the effect of cluster $C_j$ with $j\in\Lambda_i$ and $b_i\sim N(0,\sigma^2)$ is a random expert effect. Models similar to (3.1) have been successfully applied in psychometrics to describe the ratings of individuals on the items of a test or psychiatric scale. In this context, model (3.1) is known as the Rasch model and it plays an important role in the conceptualization of fundamental measurement in psychology, psychiatry, and educational testing (De Boeck and Wilson, 2004; Bond and Fox, 2007). There are clear similarities between the problem studied in this work and the measurement problem tackled in psychometrics. For instance, the clusters in our setting parallel the role of the items in a test or psychiatric scale and the ratings of the individual on these items would be equivalent to the ratings given by the experts in our setting. In addition to the intuitive meaning attached to treating clusters as fixed effects (just like items in Rasch model), it is also computationally convenient. Nonetheless, differences in the inferential target and the dimension of the parametric space imply that distinctive approaches are needed in both areas.

Parameter estimates for model (3.1) are obtained by maximizing the likelihood,

$$L(\boldsymbol{\beta},\sigma^2)=\prod_{i=1}^{n}\int_{-\infty}^{\infty}\prod_{j\in\Lambda_i}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}\;\phi(b_i|0,\sigma^2)\,db_i, \tag{3.2}$$

using, for example, a Newton-Raphson optimization algorithm, where
$\pi_{ij}=P\left(Y_{ij}=1|b_i\right)$, $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_N)$ is a vector containing all cluster effects and $\phi(b_i|0,\sigma^2)$ denotes a normal density with mean zero and variance $\sigma^2$. The integral can be approximated applying numerical procedures like Gauss-Hermite quadrature.

Using model (3.1), one can calculate the marginal probability of success for cluster $C_j$ by integrating over the distribution of the random effects:

$$P_j=P\left(Y_j=1\right)=\int\frac{\exp\left(\beta_j+b\right)}{1+\exp\left(\beta_j+b\right)}\,\phi(b|0,\sigma^2)\,db. \tag{3.3}$$

Essentially, in a first step one estimates the cluster effects $\beta_j$, after adjusting for the expert effect, by maximizing the likelihood (3.2). These estimates are then used, in a second step, to estimate the probability of success by averaging over the entire population of experts. However, the vector of fixed effects $\boldsymbol{\beta}$ in (3.2) has dimension 22,015, and the dimension of the response vector $\boldsymbol{Y}_i$ ranges from 20 to 22,015. Hence, using maximum likelihood in this scenario is not feasible with the most commonly available computing resources. The challenge is then to find a reasonable strategy to solve this high-dimensional problem when estimating the probabilities of interest.

### 3.2.1 A Permutational-Splitting Sample Procedure

Let $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$ denote the collection of ratings on the $N$ clusters, where $\mathcal{C}_j$ is a vector containing all the ratings cluster $C_j$ received. The main idea behind the procedure described in this section is the partition of the set of cluster evaluations $\mathcal{C}$ into disjoint subsets of relatively small size. As any splitting procedure, this approach raises the problem of deciding on the size of these smaller subsets. In our setting, if $N_k$ denotes the number of vectors $\mathcal{C}_j$ in the $k$ subset (where $N_1 + N_2 + \ldots, +N_S = N$ and $S$ is the total number of subsets), then one needs to determine the $N_k$'s so that model (3.1) can be fitted, with commonly available computing resources, using maximum likelihood and the information in each subset. Even though the search for appropriate $N_k$'s may produce more than one plausible choice, a sensitivity analysis could easily explore the impact of these choices on the conclusions. For instance, in our case study, very similar results were obtained with $N_k = 15$ and 30, indicating a degree of robustness with respect to this choice. In general, the choice of the subsets' cardinality may vary from one application to another. However, values of around $30 - 40$ clusters per subset seem to be a reasonable starting point. Clearly, the choice of $N_k$ automatically determines $S$ and it is possible that some subsets might have slightly more or less clusters because $S = N/N_k$ may not be a whole number. Taking these ideas into account, the following procedure is implemented:

1. **Splitting**: The set $\mathcal{C}$ is split into $S$ mutually exclusive and exhaustive subsets $\mathcal{C}^k$ $(k = 1, \ldots, S)$ with $N_k < N$ denoting the corresponding cardinality. The information in these subsets may not be independent, as ratings from the same expert may appear in more than one subset. Moreover, given that the subsets are exclusive and exhaustive, all the information needed to estimate the effect of a given cluster, say the vector of ratings $\mathcal{C}_j$, is contained in one single subset. While it is possible to include overlapping subsets into the methodology as well, this is not necessary in view of bias, etc. The most important consideration is as to whether all parameters to be estimated retain information from the partition.

2. **Estimation**: Using maximum likelihood and the information included in each $\mathcal{C}^k$, model (3.1) is fitted $S$ times. For all $k$, $N_k < N$ (typically $N_k << N$) and, consequently, the dimensions of the response and fixed-effect vectors associated with these models are now much smaller. Pooling all estimates obtained from these fittings leads to an estimate for the vector of fixed-effect parameters and $S$ estimates for the random-effect variance $\sigma^2$. Clearly, within each subset, the estimator for the inter-expert variability $\widehat{\sigma}_k^2$ uses information from only a subgroup

of all experts and, therefore, it delivers a less efficient estimate of this parameter than the estimator based on the entire data. The pooling of the subset-specific estimates should not be done mechanically and a careful analysis should be carried out to detect unusual behavior. In this regard, the procedure described in the next step may help check the stability of the parameter estimates.

3. **Permutation**: The elements of $\mathcal{C}$ are randomly permuted and steps 1 and 2 repeated $W$ times. This step is equivalent to sampling without replacement from the set of all possible partitions introduced in step 1. Consequently, instead of estimating the parameters of interest based on a single, arbitrary partition, their estimation is now based on multiple, randomly selected partitions of the set of clusters. The permutation step serves several purposes. It allows for the estimation of the parameters based on different subsamples of the same data and, hence, it makes possible to check the stability of these estimates. This may be especially relevant for the variance component, since it is estimated under different sample sizes. In addition, by combining estimates from different subsamples it produces more reliable final estimates. To capitalize on these issues, one should ideally consider a large number of permutations ($W$), our results however, indicate little gain by taking $W$ larger than 20.

4. **Estimating of the success probabilities**: Step 3 produces the set of estimates $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\sigma}^2_{kw}$, where $w = 1, \ldots, W$ and $k = 1, \ldots, S$. Subsequently, based on $\widehat{\boldsymbol{\beta}}_w$ and $\widehat{\sigma}^2_w = \frac{1}{S}\sum_{k=1}^{S}\widehat{\sigma}^2_{kw}$, estimates of the success probability of every cluster can be obtained using (3.3), with the integral computed via Monte Carlo integration by drawing $Q$ elements $b_q$ from $N(0, \widehat{\sigma}^2_w)$. It is important to note that, unlike the $\widehat{\sigma}^2_{kw}$ that only uses information from the experts in the $k$th subset, $\widehat{\sigma}^2_w$ is based on information from all experts and, hence, it offers a better assessment of the inter-expert variability. Eventually, the probability of success for cluster $C_j$ can be estimated as

$$\widehat{P}_j = \frac{1}{W}\sum_{w=1}^{W}\widehat{P}_{wj}, \quad \text{where} \quad \widehat{P}_{wj} = \widehat{P}_w\left(Y_j = 1\right) = \frac{1}{Q}\sum_{q=1}^{Q}\frac{\exp\left(\widehat{\beta}_{wj} + b_q\right)}{1 + \exp\left(\widehat{\beta}_{wj} + b_q\right)}.$$

Similarly,

$$\widehat{\beta}_j = \frac{1}{W}\sum_{w=1}^{W}\widehat{\beta}_{wj}, \qquad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{W}\sum_{w=1}^{W}\widehat{\sigma}^2_w.$$

One may heuristically argue that step (3) also ensures that final estimates of the cluster effects are similar to those obtained when maximum likelihood is used

with the whole data. Indeed, let $\widehat{\beta}_{wj}$ denote again the maximum likelihood estimators for the effect of cluster $C_j$ computed in each of the $W$ permutations and $\widehat{\beta}_{Nj}$ the maximum likelihood estimator based on the entire set of $N$ clusters. Further, consider the expression $\widehat{\beta}_{wj} = \widehat{\beta}_{Nj} + e_{wj}$, where $e_{wj}$ is the random component by which $\widehat{\beta}_{wj}$ differs from $\widehat{\beta}_{Nj}$. Given that maximum likelihood estimators are asymptotically unbiased, one has $E(e_{wj}) \approx 0$ and extensions of the law of large numbers for correlated, not identically distributed random variables may suggest that, under certain assumptions, for a sufficiently large $W$ (Newman, 1984; Birkel, 1992)

$$\widehat{\beta}_j = \frac{1}{W}\sum_{w=1}^{W}\widehat{\beta}_{wj} = \widehat{\beta}_{Nj} + \frac{1}{W}\sum_{w=1}^{W} e_{wj} \approx \widehat{\beta}_{Nj}.$$

Similar arguments could be put forward for the variance component and the success probabilities as well. The findings of the simulation study presented in Section 3.4 support these heuristic results.

5. **Confidence interval for the success probabilities**: To construct a confidence interval for the success probability of cluster $C_j$, we consider the results from one of the $W$ permutations described in step 3. To simplify notation, we omit the subscript $w$ in the following equations, but these calculations are meant to be done for each of the $W$ permutations.

   If $\mathcal{C}^k$ denotes the unique subset of $\mathcal{C}$ containing $\mathcal{C}_j$, then fitting model (3.1) to $\mathcal{C}^k$ produces the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_j = (\widehat{\beta}_j, \widehat{\sigma}_k^2)'$. Classical likelihood theory guarantees that, asymptotically, $\widehat{\boldsymbol{\theta}}_j \sim N(\boldsymbol{\theta}_j, \boldsymbol{\Sigma})$, where a consistent estimator of the $2\times 2$ matrix $\boldsymbol{\Sigma}$ can be constructed using the Hessian matrix obtained upon fitting the model. Even though the estimator $\widehat{\sigma}_k^2$ is not efficient, its use is necessary in this case to directly apply asymptotic results from maximum likelihood theory.

   The success probability $P_j$ is a function of $\boldsymbol{\theta}_j$, such that, if one defines $\gamma_j = \log\{P_j/(1-P_j)\}$, then the delta method leads to $\widehat{\gamma}_j \sim N\left(\gamma_j, \sigma_\gamma^2\right)$ asymptotically, where $\widehat{\gamma}_j = \log\left\{\widehat{P}_j/(1-\widehat{P}_j)\right\}$ and

$$\sigma_\gamma^2 = \left(\frac{\partial\gamma_j}{\partial\boldsymbol{\theta}_j}\right)\boldsymbol{\Sigma}\left(\frac{\partial\gamma_j}{\partial\boldsymbol{\theta}_j}\right)',$$

$$\frac{\partial\gamma_j}{\partial\boldsymbol{\theta}_j} = \frac{1}{P_j(1-P_j)}\frac{\partial P_j}{\partial\boldsymbol{\theta}_j},$$

with

$$\frac{\partial P_j}{\partial \beta_j} = \int \frac{\exp(\beta_j + b)}{\{1 + \exp(\beta_j + b)\}^2} \, \phi(b|0, \sigma_k^2) \, db,$$

$$\frac{\partial P_j}{\partial \sigma_k^2} = \int \frac{\exp(\beta_j + b)}{1 + \exp(\beta_j + b)} \, \frac{b^2 - \sigma_k^2}{2\sigma_k^4} \, \phi(b|0, \sigma_k^2) \, db.$$

The necessary estimates can be obtained from plugging $\widehat{\boldsymbol{\theta}}_j$ into the corresponding expressions and using Monte Carlo integration as previously described. Finally, an asymptotic 95% confidence interval for $P_j$ is given by

$$CI_{P_j} = \frac{\exp(\widehat{\gamma}_j \pm 1.96 \cdot \widehat{\sigma}_\gamma)}{1 + \exp(\widehat{\gamma}_j \pm 1.96 \cdot \widehat{\sigma}_\gamma)}.$$

The overall confidence interval follows from averaging the lower and upper bounds of all confidence intervals from the $W$ partitions. If information is not uniformly divided over subsamples, then weighted averages rather than averages need to be used. In principle, one should adjust the coverage probabilities using, for example, the Bonferroni correction when constructing these intervals. If the overall coverage probability for the entire family of confidence intervals is 95%, then it is easy to show that the final average interval will have a coverage probability of at least 95%. This implies construction of confidence intervals to the level of $(1 - 0.05/W)$ for $P_j$ in each permutation, which are likely to be too wide for useful inference. In Section 3.4, we study the performance of this interval via simulation without using any correction, and the results confirm that in many practical situations this simpler approach may work well.

## 3.3 Data Analysis

### 3.3.1 Unweighted Analysis

The procedure introduced in Section 3.2 was applied to the data described in Section 2.1, using $N_k = 30$, $Q = 10,000$, $S = 734$ and $W = 20$. Table 3.1 gives the results for the 20 top-ranked clusters, i.e., the clusters with the highest estimated probability of success. All clusters in the table have an estimated probability larger than 60%, and the top 3 have probability of success around 75%. The observed probabilities (proportion of ones for each cluster), are substantially different from the model estimated probabilities for some clusters. Importantly, the proportions completely ignore the correlation between ratings from the same expert. Therefore, they do not correct for

the fact that some experts may tend to give higher/lower ratings than others and may lead to biased estimates for clusters that are mostly evaluated by definite/skeptical experts. In addition, the results also indicate a high heterogeneity among experts, with estimated variance

$$\widehat{\sigma}^2 = \frac{1}{W}\sum_{w=1}^{W}\widehat{\sigma}_w^2 \approx 10.$$

On the one hand, this large variance may indicate the need for selecting experts from a more uniform population by defining, for example, more stringent selection criteria. On the other hand, more stringent selection criteria may conflict with having experts that represent an appropriately broad range of expert opinion. Finding a balance between these two considerations is very important to guarantee the overall quality of the study. In general, if substantial heterogeneity among experts is encountered, then additional investigations should try to determine the source before further actions are taken.

The general behavior of the estimated probabilities of success is displayed in Figure 3.1. Visibly, most clusters have a quite low probability of success, with the median around 26%, and 75% of the clusters have an estimated probability of success smaller than 40%. About 100 clusters are unanimously not recommended, as evidenced by the peak at zero probability. This is in line with the observed data, given that none of them got a positive recommendation despite their number of evaluations ranging between 11 and 23. Another conspicuous group represents clusters that had only 1–3 positive evaluations and, as expected, produced low estimated probabilities of success raging between 0.08–0.1.

The interpretation of these probabilities will frequently be subject-specific. Taking into account the economic cost associated with the development of these clusters, the time frame required to develop them, and the potential social and economic gains that they may bring, researchers can define the minimum probability of success that may justify further study.

The analysis of the confidence intervals also offers some important insight. First, although moderately wide, the confidence intervals still allow for useful inferences. Actually, the large inter-expert heterogeneity may hint at possible measures to increase precision in future studies. Second, using the lower bound of the confidence intervals to rank the clusters, instead of the point estimate of the probability of success, may yield different results. By this criterion, cluster 265,222, ranked eighth by the point estimate, would become the second most promising candidate. Clearly, some more fundamental, substantive considerations may be needed to complement the information in Table 3.1 during the decision making process.

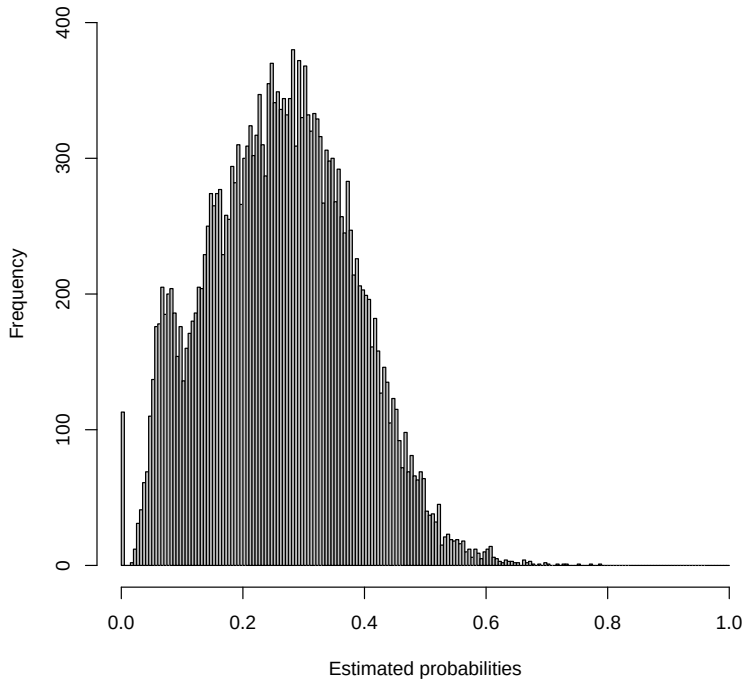


Figure 3.1: *Distribution of estimated probabilities of success.*

As a sensitivity analysis we also considered $N_k = 15$, $W = 20$, $S = 1468$. The results obtained appear in the columns labeled "unweighted" in Table 3.2. Clearly, the differences with the original analysis are negligible except for the $\widehat{\sigma}^2$, resulting into slight changes in the rankings.

### 3.3.2 Weighted Analysis

An important issue discussed in Section 2.1.2 was the differences encountered in the number of clusters evaluated by the experts. One may wonder whether experts who evaluated a large number of clusters gave as careful consideration to each cluster as those who evaluated only a few. Importantly, the model-based approach introduced in Section 3.2 can take into account these differences by carrying out a weighted analysis

which maximizes the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} W_i \int_{-\infty}^{\infty} \prod_{j \in \Lambda_i} \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \, \phi(b_i | 0, \sigma^2) \, db_i, \tag{3.4}$$

where $W_i = N/|\Lambda_i|$ and $|\Lambda_i|$ denotes the cardinality of $\Lambda_i$. Practically, a weighted analysis using the SAS procedure NLMIXED, implies replication of each response vector by $W_i$, resulting into a pseudo-dataset with larger sample size than in the unweighted analysis. Using partitions with $N_k = 30$ was rather challenging and, consequently, the weighted analysis was carried out with $N_k = 15$ and this was adopted for all the other analyses. The main results are displayed in Table 3.2.

Interestingly, some important differences emerge from the two approaches. For instance, the top-ranked cluster in the unweighted analysis received rank 2 in the weighted approach. Some differences are even more dramatic; for example, the fourth cluster in the unweighted analysis received rank 620 in the weighted approach. Clearly, a very careful and thoughtful discussion of these differences will be needed during the decision making process. In addition, these results also point out the importance of a careful design of the study and may suggest to introduce changes in the design to avoid large differences in the number of clusters evaluated by the experts. The cluster ranked 20th is not in Table 3.1, probably due to the change in $\widehat{\sigma}^2$.

Fitting model (3.1) to the entire data set using maximum likelihood was unfeasible in this case study. Therefore, all previous conclusions were derived by implementing the procedure described in Section 3.2. One may wonder how the previous procedure would compare with maximum likelihood when the latter is tractable. In the next section we investigate this important issue via simulation.

## 3.4 Simulation Study

The simulations were designed to mimic the main characteristics encountered in the case study. Two hundred datasets were generated, with the following parameters held constant in all datasets: (1) Number of clusters, $N = 50$, chosen to ensure tractability of maximum likelihood estimation for the whole data, (2) number of experts, $n = 147$, and (3) a set of 50 values assigned to parameters characterizing the cluster-effects ($\beta_j$), which was sampled from a $N(-2, 2)$. Factors varying across the datasets were: the number of ratings per expert, ($N_i$), independently sampled from Poisson(25) and restricted to the range of 8 to 50 and a set of of 147 random-effects ($b_i$), independently sampled from $N(0, 12.25)$. They were varied by using a dataset specific seed in generating $N_i$ and $b_i$. It is only logical that $b_i$ varies from one dataset to

another as a way of creating different random samples from the experts' population and $N_i$ needs to simultaneously vary with $b_i$. Each expert rated all the assigned clusters and the assignment of number of ratings was random. This aspect is different from the case study, and it was done to avoid complications arising from the design as discussed in Section 2.1.2, so that the simulations investigate the performance of the procedure without being confounded by the validity of the model used. Based on these values, the probability that $i^{th}$ expert rates the $j^{th}$ cluster $P_{ij} = P(Y_{ij} = 1|b_i)$, was computed using model (3.1) and the response $Y_{ij} \sim$ Bernoulli$(P_{ij})$. Finally, model (3.1) was fitted using full maximum likelihood and the procedure introduced in Section 3.2 and their corresponding probabilities of success, given by (3.3), were compared. Parameters used in split-permutation procedure were: $N_s = 5$, $W = 20$, $Q = 10,000$ and $S = 10$.

The main results of the simulation study for the top 20 clusters (those with the highest true probability of success) are summarized in Tables 3.3 and 3.4. Regarding the point estimates of the cluster-effect, Table 3.3 clearly shows that the proposed procedure performs as well as maximum likelihood. Figure 3.2 shows that this is true for most of the clusters as the relative differences between the true values and each of the methods' estimates are close to zero. Full maximum likelihood cluster-effect estimates for clusters 14, 27, and 30, have noticeably larger relative bias than their split-procedure counterparts. An inspection of the estimates of these clusters from the 20 permutations, clearly shows that even estimates from the sub-samples, were biased, though the pooled estimates have reduced bias. This underscores the observation in Step 2 that, the mean of random errors arising from estimating the cluster-effects in a model with less clusters than the total would approximates to zero and also emphasizes the importance of the permutation step.

Further scrutiny of the estimated success probabilities in (Table 3.4), rubber stamps the similarity in performance between the two methods. Importantly, the point estimates are very close to the true values in both cases, and the coverage of the confidence intervals are around 95% when full likelihood is used, and is slightly higher for the new procedure, a possible consequence of using a less efficient $\sigma^2_{sw}$. The percentage of confidence intervals that were entirely below the true value and those entirely above are similar for both methods. Nevertheless, faced with the possibility of not being able to analyze the data, a little loss in precision seems a reasonable price to pay. In spite of the small differences, fitting model (3.1) using full maximum likelihood and the procedure introduced in Section 3.2 basically yield the same results.

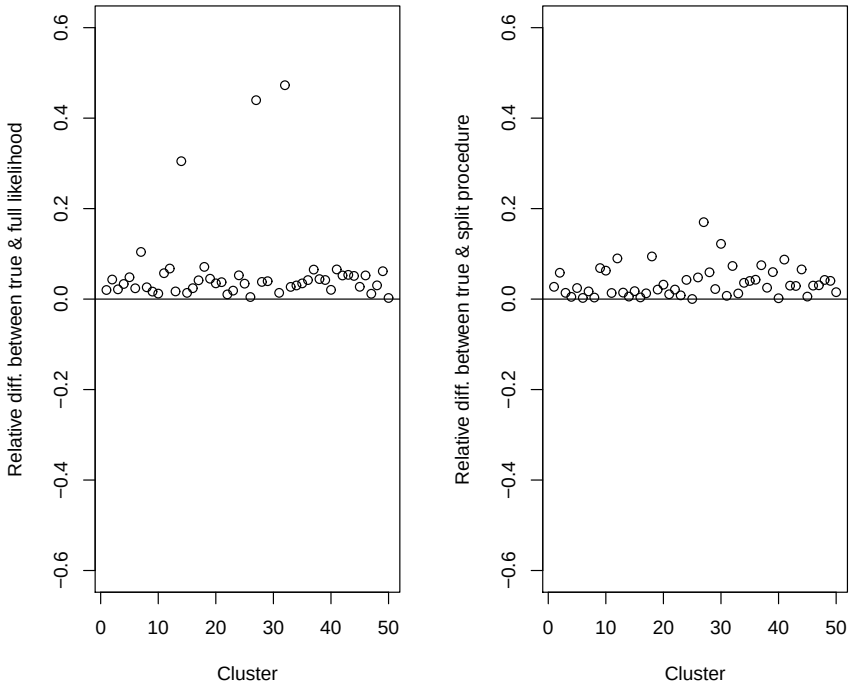


Figure 3.2: *Relative difference between the true values and the estimates obtained from maximum likelihood (mle),* $\frac{\beta_j - \widehat{\beta}_j}{\beta_j} mle$ *(left) and the permutational-splitting procedure(split)* $\frac{\beta_j - \widehat{\beta}_j}{\beta_j} split$ *(right).*

## 3.5 Discussion

In our quest to quantify expert opinion on the potential of clusters of chemical compounds, we have introduced a *permutational-splitting sample procedure*. A combination of maximum likelihood estimation, re-sampling, and Monte Carlo methods produced parameters estimates and confidence intervals, comparable to those obtained from full maximum likelihood. Loss in precision with the permutational-splitting sample procedure, apparent in wider confidence intervals is anticipated, since the procedure splits the data into dependent sub-samples, resulting into a less efficient random-effect variance estimate.

The model used for the statistical analysis and the conclusions derived from it rest on a number of assumptions, like the distribution of the expert-specific effect $b_i$. Although the normality assumption for the random effects is standard in most software packages, in principle, it would be possible to consider other random-effect distributions. For instance, using probability integral transformations in the SAS procedure NLMIXED, other distribution could be fitted as well, but obtaining convergence is much more challenging with these models Nelson *et al.*.

One could also conceive extending the model by letting the rater effects vary across cluster. However, this extension will dramatically increase the dimension of the vector of random effects, aggravating the already challenging numerical problems. In general, the successful application of the Rash model in psychometrics to tackle problems similar to the one considered here, makes us believe that, although it cannot be formally proven, model (3.1) may offer a feasible and reliable way to estimate the success probabilities of interest.

Obviously, more simulation studies and applications to real problems will shed light on the potential and limitations of the model and fitting procedure proposed in the present work. Importantly, their application is possible with commonly available software and a simulated data set with the corresponding SAS code for the analysis can be freely downloaded from http://www.ibiostat.be/software/.

Even though it was not the focus of the present work, it is clear that the design of the study is another important element to guarantee the validity of the results. Optimal designs are a class of experimental designs that are optimal with respect to some statistical criterion (Berger and Wong, 2009). For instance, one may aim to select the number of experts, the number of clusters assigned to the experts and the assignment mechanism to maximize precision when estimating the probabilities of success. In principle, it seems intuitively desirable for each cluster to be evaluated by the same number of experts and for each pair of experts to have a reasonable number

of clusters in common. However, more research will be needed to clarify these issues and establish the best possible design for this type of studies.

Table 3.1: *Top* 20 *clusters (ID) with highest estimated probability of success: Estimated cluster-effect ($\hat{\beta}_j$), Estimated/Observed success probabilities (proportion of ones for each cluster) and confidence interval limits.*

| | | probability | | 95% C.I. | |
|---|---|---|---|---|---|
| ID | $\hat{\beta}_j$ | estimated | observed | lower | upper |
| 295061 | 3.07 | 0.80 | 0.82 | 0.58 | 0.92 |
| 296535 | 2.51 | 0.76 | 0.81 | 0.51 | 0.90 |
| 84163 | 2.40 | 0.75 | 0.78 | 0.48 | 0.90 |
| 313914 | 2.30 | 0.74 | 0.80 | 0.39 | 0.93 |
| 265441 | 2.16 | 0.72 | 0.69 | 0.50 | 0.87 |
| 296443 | 2.09 | 0.72 | 0.62 | 0.52 | 0.86 |
| 277774 | 2.01 | 0.71 | 0.71 | 0.49 | 0.86 |
| 265222 | 1.96 | 0.71 | 0.70 | 0.53 | 0.84 |
| 178994 | 1.84 | 0.69 | 0.73 | 0.50 | 0.84 |
| 462994 | 1.73 | 0.69 | 0.69 | 0.44 | 0.86 |
| 292579 | 1.76 | 0.69 | 0.75 | 0.45 | 0.84 |
| 296560 | 1.71 | 0.68 | 0.72 | 0.47 | 0.83 |
| 277619 | 1.67 | 0.68 | 0.63 | 0.47 | 0.83 |
| 315928 | 1.67 | 0.68 | 0.75 | 0.47 | 0.84 |
| 296427 | 1.69 | 0.68 | 0.78 | 0.35 | 0.91 |
| 263047 | 1.60 | 0.68 | 0.76 | 0.45 | 0.84 |
| 333529 | 1.62 | 0.67 | 0.80 | 0.45 | 0.84 |
| 292805 | 1.52 | 0.67 | 0.72 | 0.43 | 0.85 |
| 178828 | 1.43 | 0.66 | 0.72 | 0.43 | 0.83 |
| 265229 | 1.39 | 0.65 | 0.65 | 0.47 | 0.80 |
| $\widehat{\sigma}^2$ | 10.279 | | | | |

Table 3.2: *Estimates for the fixed effects and probabilities of success obtained from the weighted and unweighted analyses for previously reported top* 20 *clusters* $\hat{\beta}_{weighted}$ *and* $\hat{\beta}_{unweighted}$ *are the estimated cluster-effect with the ranks in brackets, and* $\widehat{prob}_{weighted}$ *and* $\widehat{prob}_{unweighted}$ *are the corresponding probabilities of success.*

| ID | $\hat{\beta}_{\text{weighted}}$ | $\hat{\beta}_{\text{unweighted}}$ | $\widehat{\text{prob}}_{\text{weighted}}$ | | $\widehat{\text{prob}}_{\text{unweighted}}$ | |
|---|---|---|---|---|---|---|
| 295061 | 3.86 | 3.33 | 0.90 | ( 2) | 0.80 | ( 1) |
| 296535 | 1.99 | 2.71 | 0.74 | ( 54) | 0.76 | ( 2) |
| 84163 | 0.86 | 2.42 | 0.61 | ( 376) | 0.73 | ( 3) |
| 296443 | 0.54 | 2.41 | 0.57 | ( 620) | 0.73 | ( 4) |
| 313914 | 3.79 | 2.37 | 0.89 | ( 3) | 0.73 | ( 5) |
| 265222 | 0.56 | 2.40 | 0.57 | ( 653) | 0.73 | ( 6) |
| 333529 | 1.85 | 1.99 | 0.73 | ( 67) | 0.69 | ( 7) |
| 296560 | 1.26 | 1.91 | 0.66 | ( 198) | 0.69 | ( 8) |
| 178994 | 2.25 | 1.91 | 0.77 | ( 28) | 0.69 | ( 9) |
| 265441 | 1.22 | 1.94 | 0.66 | ( 211) | 0.69 | (10) |
| 277774 | 2.26 | 1.87 | 0.77 | ( 29) | 0.69 | (11) |
| 292579 | 2.69 | 1.91 | 0.81 | ( 10) | 0.69 | (12) |
| 315928 | 1.18 | 1.87 | 0.65 | ( 233) | 0.68 | (13) |
| 277619 | -0.63 | 1.74 | 0.42 | (3165) | 0.67 | (14) |
| 263047 | 3.85 | 1.78 | 0.90 | ( 1) | 0.67 | (15) |
| 296427 | 2.70 | 1.65 | 0.81 | ( 12) | 0.67 | (16) |
| 292805 | 1.00 | 1.60 | 0.63 | ( 313) | 0.66 | (17) |
| 178828 | 2.26 | 1.52 | 0.77 | ( 27) | 0.66 | (18) |
| 462994 | 1.31 | 1.46 | 0.67 | ( 183) | 0.65 | (19) |
| 159643 | 1.93 | 1.50 | 0.74 | ( 55) | 0.65 | (20) |
| $\widehat{\sigma}^2$ | 3.19 | 15.80 | | | | |

Table 3.3: *True values and average parameter estimates for the top* 20 *clusters (ID), estimated from full likelihood (likelihood) and the permutational-splitting procedure (procedure).*

| | $\beta_j$ | | |
|---|---|---|---|
| ID | true | likelihood | procedure |
| 3 | 2.33 | 2.38 | 2.36 |
| 1 | 1.60 | 1.63 | 1.65 |
| 33 | 1.52 | 1.56 | 1.54 |
| 47 | 1.43 | 1.45 | 1.48 |
| 50 | 1.04 | 1.03 | 1.05 |
| 27 | 0.13 | 0.07 | 0.11 |
| 30 | 0.06 | 0.01 | 0.05 |
| 32 | 0.06 | 0.03 | 0.06 |
| 14 | -0.11 | -0.14 | -0.11 |
| 7 | -0.30 | -0.33 | -0.29 |
| 9 | -0.49 | -0.50 | -0.46 |
| 48 | -0.63 | -0.65 | -0.61 |
| 10 | -0.71 | -0.70 | -0.66 |
| 21 | -0.97 | -1.00 | -0.98 |
| 11 | -1.12 | -1.19 | -1.14 |
| 26 | -1.13 | -1.12 | -1.07 |
| 15 | -1.32 | -1.33 | -1.29 |
| 13 | -1.40 | -1.42 | -1.38 |
| 4 | -1.42 | -1.47 | -1.42 |
| 42 | -1.61 | -1.69 | -1.66 |
| $\widehat{\sigma}^2$ | 12.25 | 12.96 | 12.74 |

Table 3.4: *Average estimated success probabilities for top* 20 *clusters (ID), using full likelihood (lik) and the permutational-splitting procedure (proc), percentage of coverage of the confidence intervals (coverage %), percentage of times the true value was less than lower confidence limit, (non-cov(below) %), and percentage of times the true value was greater than upper confidence limit, (non-cov(below)*

| | | probability of success | | | coverage % | | non-cov(below)% | | non-cov(above)% | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rank | ID | true | lik | proc | lik | pro | lik | proc | lik | proc |
| 1 | 3 | 0.72 | 0.72 | 0.73 | 0.94 | 0.95 | 0.02 | 0.02 | 0.05 | 0.04 |
| 2 | 1 | 0.66 | 0.66 | 0.66 | 0.95 | 0.96 | 0.03 | 0.02 | 0.03 | 0.03 |
| 3 | 33 | 0.65 | 0.65 | 0.65 | 0.98 | 0.97 | 0.01 | 0.01 | 0.02 | 0.02 |
| 4 | 47 | 0.64 | 0.64 | 0.65 | 0.96 | 0.96 | 0.02 | 0.02 | 0.02 | 0.02 |
| 5 | 50 | 0.60 | 0.60 | 0.61 | 0.96 | 0.96 | 0.02 | 0.02 | 0.03 | 0.01 |
| 6 | 27 | 0.51 | 0.51 | 0.51 | 0.96 | 0.96 | 0.02 | 0.02 | 0.03 | 0.02 |
| 7 | 30 | 0.51 | 0.50 | 0.51 | 0.93 | 0.94 | 0.03 | 0.02 | 0.04 | 0.03 |
| 8 | 32 | 0.51 | 0.50 | 0.51 | 0.94 | 0.96 | 0.04 | 0.02 | 0.03 | 0.01 |
| 9 | 14 | 0.49 | 0.49 | 0.49 | 0.97 | 0.96 | 0.01 | 0.01 | 0.03 | 0.03 |
| 10 | 7 | 0.47 | 0.47 | 0.47 | 0.94 | 0.96 | 0.01 | 0.02 | 0.05 | 0.02 |
| 11 | 9 | 0.45 | 0.45 | 0.45 | 0.97 | 0.96 | 0.02 | 0.02 | 0.02 | 0.02 |
| 12 | 48 | 0.44 | 0.44 | 0.44 | 0.96 | 0.96 | 0.03 | 0.03 | 0.01 | 0.01 |
| 13 | 10 | 0.43 | 0.43 | 0.43 | 0.92 | 0.95 | 0.04 | 0.03 | 0.05 | 0.03 |
| 14 | 21 | 0.40 | 0.40 | 0.40 | 0.97 | 0.97 | 0.02 | 0.02 | 0.01 | 0.01 |
| 15 | 11 | 0.39 | 0.38 | 0.39 | 0.95 | 0.95 | 0.03 | 0.03 | 0.03 | 0.02 |
| 16 | 26 | 0.39 | 0.39 | 0.39 | 0.94 | 0.95 | 0.04 | 0.04 | 0.02 | 0.01 |
| 17 | 15 | 0.37 | 0.37 | 0.37 | 0.96 | 0.97 | 0.03 | 0.02 | 0.01 | 0.01 |
| 18 | 13 | 0.36 | 0.36 | 0.36 | 0.95 | 0.96 | 0.04 | 0.03 | 0.02 | 0.02 |
| 19 | 4 | 0.36 | 0.36 | 0.36 | 0.94 | 0.95 | 0.03 | 0.03 | 0.04 | 0.02 |
| 20 | 42 | 0.34 | 0.34 | 0.34 | 0.95 | 0.97 | 0.04 | 0.02 | 0.02 | 0.01 |

# Chapter 4

# Impact of Selection Bias on the Qualitative Assessment of Clusters of Chemical Compounds

The steady advance taking place in fields like genetics and molecular biology, is dramatically increasing our capacity to obtain new drugs. Nevertheless, developing a chemical compound into an effective drug is often an expensive and lengthy process. As a result, one needs to carefully evaluate the amount of evidence that supports the potential of certain compounds before investing more resources into it (Alonso *et al.*, 2008). Chapter 3, noted that in addition to in-house, chemical compounds libraries, pharmaceutical companies, occasionally acquire such potential compounds from third party vendors. Also discussed was a new technique by Hack *et al.* (2011) to aid the selection of appropriate clusters of compounds to acquire. In Section 2.1.2, several statistical challenges associated with this technique are discussed and Chapter 3 tackles the high-dimensional data challenge. This chapter investigates problems arising from the less restricted assignment of clusters to the experts.

## 4.1 Introduction

The large number of clusters of chemical compounds that may be considered for acquisition, implies that a selection procedure, by which every expert chooses or gets assigned to a number of clusters for evaluation, needs to be implemented. The approach discussed in Section 2.1, allows the experts to decide on the number of clusters they want to evaluate depending on their schedules. In the present work, we argue that such a procedure may lead to serious selection bias that can jeopardize the entire evaluation process. Two possible strategies to avoid the previous problem are: (i) to compel every expert to evaluate all clusters and (ii) to assign a single subset of the clusters to experts randomly and compel them to finish their quota. Strategy (i) may be practically infeasible, given the exorbitant number of candidates one frequently is confronted with in this type of studies. Implementing strategy (ii) may lead to some logistic difficulties, but it arguably is the most reasonable and reliable option to avoid bias and simplify the posterior analysis of the data, we strongly advocate (ii).

Problems that come with selection bias, as well as their possible correction, have been documented in many fields (Horwitz and Feinstein, 1978; Hernán, Hernández-Diaz, and Robins, 2004; Geneletti, Richardson, and Best, 2009). Geneletti *et al.* (2011) noted that the crucial factor to determine the most appropriate bias correction method is the underlying cause of bias. This is apparent in the methods available in the literature, given that most of them are tailored towards a specific form of bias origin (Torner *et al.*, 2010; Heckman, 1979; Puhani, 2000; Lee and Marsh, 2000; Baser *et al.*, 2003; Jüni and Egger, 2005). A key similarity in the methods discussed by some of these authors is the formulation of separate models for the outcome and the selection process. Typically untestable assumptions are associated with these models, simply because the outcomes of subjects that were not selected are never known.

Using theoretical elements and simulations, we show that, in the presence of selection bias, the probability of success for every cluster can be estimated only by making strong and untestable assumptions. However, an upper bound for this probability may be obtained under a weaker condition of monotonicity.

### 4.1.1 Naive Estimation of Probabilities

In Chapter 3, success probabilities were estimated without taking into account the assignment of clusters as follows: denote the vector of ratings associated with expert $i$ by $\boldsymbol{Y}_i = (Y_{ij})_{j\in\Lambda_i}$, where $\Lambda_i$ is the subset of all clusters evaluated by the $i$th expert

and $i = 1, \ldots, n$. A natural choice to model these data is the logistic-normal model

$$\text{logit}\left[P\left(Y_{ij} = 1|b_i\right)\right] = \beta_j + b_i, \tag{4.1}$$

where $\beta_j$ is a fixed parameter characterizing the effect of cluster $C_j$ with $j \in \Lambda_i$ and $b_i \sim N(0, \sigma^2)$ is a random expert effect. Based on model (4.1), the marginal probability of success for cluster $C_j$ was obtained as:

$$P_j = P\left(Y_j = 1\right) = \int \frac{\exp\left(\beta_j + b\right)}{1 + \exp\left(\beta_j + b\right)} \phi(b|0, \sigma^2)\, db. \tag{4.2}$$

where $\phi(b|0, \sigma^2)$ denotes a normal density with mean zero and variance $\sigma^2$. Results for the top 20 ranked clusters, i.e., the clusters with the highest estimated probability of success are given in the first part of Table 4.3 (under the 'Naive' columns). The median estimated probability of success for all clusters was around 18%, rather a low value, and 75% of the clusters had estimated probabilities of success smaller than 29%. However, at the top 20, all clusters had an estimated probability larger than 60% and those in the top 3 had probabilities of success around 90%.

In addition, we also found a lot of heterogeneity between experts with an estimated variance $\widehat{\sigma}^2 \approx 16$ (where$N_s = 15$). As pointed out in Chapter 3, this high inter-expert variability may have an impact on the precision of the estimates and, consequently, it may hint on the need to select experts from a more uniform population by defining, for example, more stringent selection criteria.

Finally and taking into account practical considerations like the economic cost associated with the development of these clusters, the time frame required for such a development and the social and economical gains that these clusters may bring, researchers could define the minimum probability of success that may justify inclusion into the library.

A limitation of the previous study was the varying numbers of clusters which every expert evaluated. This raises concerns about the possible presence of selection bias. In the next section, this important issue is studied in more detail. The problem of high-dimensional data is suppressed because it was addressed in Chapter 3.

## 4.2 Selection Bias

Let $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iN})$ denote the vector containing the selection-indicators for expert $i$, where $X_{ij} = 1$ if expert $i$ evaluates cluster $j$ and 0 otherwise. The probability that expert $i$ would rate cluster $j$ as 1, given that he actually evaluates it, can be

conceptualized as

$$P\left(Y_{ij}=1|X_{ij}=1,a_i,b_i\right)=\frac{P\left(Y_{ij}=1,X_{ij}=1|a_i,b_i\right)}{P\left(X_{ij}=1|a_i,b_i\right)}, \tag{4.3}$$

where $(a_i, b_i)$ is a vector of expert-specific random effects, assumed to follow a bivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$. We say that there is selection bias in the rating process if

$$P\left(Y_{ij}=y_{ij}|X_{ij}=1,a_i,b_i\right)\neq P\left(Y_{ij}=y_{ij}|X_{ij}=0,a_i,b_i\right).$$

Furthermore, the following conditional independence assumption will play an important role in some of the subsequent developments

$$P\left(Y_{ij}=y_{ij},X_{ij}=x_{ij}|a_i,b_i\right)=P\left(Y_{ij}=y_{ij}|b_i\right)P\left(X_{ij}=x_{ij}|a_i\right), \tag{4.4}$$

for all $i, j$. Essentially, (4.4) states that for every expert the rating and selection procedures are independent and governed by different, although possibly correlated, random effects. Some important scenarios covered by (4.4) are the ones described as strategy (i) and (ii) in Section 4.1. Indeed, in strategy (i) all experts are compelled to evaluate all clusters and, therefore, $P\left(X_{ij}=1|a_i\right)=1$ for all $i$, $j$. Moreover, in strategy (ii) the possible dependence between $Y_{ij}$ and $X_{ij}$ is broken by the random allocation and in that case typically $P\left(X_{ij}=1|a_i\right)=P\left(X_{ij}=1\right)$. Under (4.4), expression (4.3) can be rewritten as

$$P\left(Y_{ij}=1|X_{ij}=1,a_i,b_i\right)=P\left(Y_{ij}=1|X_{ij}=1,b_i\right)=P\left(Y_{ij}=1|b_i\right). \tag{4.5}$$

Model (4.1), used in Section 4.1.1 to quantify the success probabilities, basically tries to characterize $P\left(Y_{ij}=1|b_i\right)$ and, hence, it is valid if the conditional independence assumption holds. Some comments are in place. Note first that, on the one hand, $P\left(Y_{ij}=1|b_i\right)$ quantifies the chance that expert $i$ will rate cluster $j$ as 1, irrespective of whether he actually evaluates the cluster or not. Thus, it is a marginal probability that does not depend on the selection process. On the other hand, $P\left(Y_{ij}=1|X_{ij}=1,b_i\right)$ describes the chance that expert $i$ will rate cluster $j$ as 1 given that he evaluates it and, in general, it might differ from $P\left(Y_{ij}=1|X_{ij}=0,b_i\right)$. Actually, in the most general scenario, the potential of cluster $j$ can be quantified as

$$P\left(Y_j=1\right)=\int\int P\left(Y_{ij}=1|a_i,b_i\right)\,\phi(a_i,b_i|\mathbf{0},\mathbf{\Sigma})\,da_i db_i, \tag{4.6}$$

where $\phi(\cdot|\mathbf{0}, \mathbf{\Sigma})$ denotes a bivariate normal density with mean zero and covariance matrix $\mathbf{\Sigma}$ and

$$\begin{aligned} P(Y_{ij} = 1|a_i, b_i) =& E_X \left[ P(Y_{ij} = 1|X_{ij} = x_{ij}, a_i, b_i) \right] \\ =& P(Y_{ij} = 1|X_{ij} = 1, a_i, b_i) P(X_{ij} = 1|a_i, b_i) \\ &+ P(Y_{ij} = 1|X_{ij} = 0, a_i, b_i) P(X_{ij} = 0|a_i, b_i). \end{aligned} \tag{4.7}$$

This expression is very insightful. Note first that we have information about how the experts rated the clusters they evaluated and, therefore, $P(Y_{ij} = 1|X_{ij} = 1, a_i, b_i)$ can be estimated from the data. We also have information about which clusters every expert evaluated and we could use this information to estimate $P(X_{ij} = 1|a_i, b_i)$. The critical term in (4.7) is $P(Y_{ij} = 1|X_{ij} = 0, a_i, b_i)$. In fact, the event $\{Y_{ij} = y_{ij}|X_{ij} = 0, a_i, b_i\}$ is counterfactual and we do not have information about how the experts would have rated a cluster they did not evaluate if, contrary to fact, they had evaluated it. As a result, this probability is not identifiable from the data without additional assumptions.

The previous discussion illustrates that in the most general case computing (4.6) requires: (1) to explicitly model $P(X_{ij} = 1|a_i, b_i)$ and (2) to make untestable assumptions about the *counterfactual* probabilities $P(Y_{ij} = 1|X_{ij} = 0, a_i, b_i)$. A reasonable such assumption in many situations may be the following monotonicity condition

$$P(Y_{ij} = 1|X_{ij} = 0, a_i, b_i) \leq P(Y_{ij} = 1|X_{ij} = 1, a_i, b_i).$$

That may be the case, for instance, if experts choose to evaluate those clusters that they find more promising or interesting. The previous inequality implies that

$$P(Y_{ij} = 1|a_i, b_i) \leq P(Y_{ij} = 1|X_{ij} = 1, a_i, b_i)$$

and, hence, one could use the data to provide an upper bound for (4.6). This upper bound suggests that in many applications discarding those clusters with a small estimated probability of success may be reasonable, even if selection bias is present. Nonetheless, one should be cautious when interpreting a large probability of success if selection bias is suspected.

### 4.2.1 How Ignorable is the Selection Procedure in the Absence of Selection Bias?

It is clear from the previous discussion that, in the presence of selection bias, one needs to explicitly model the selection mechanism to compute (4.6). Nevertheless, the

preceding arguments do not fully clarify whether the selection procedure can be safely ignored when selection bias is not present. In what follows we will assume conditional independence as a natural way to avoid selection bias and study the ignorability of the selection process in some detail, but first we need to extend notation. Let $P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j)$ and $P(Y_{ij} = y_{ij}|X_{ij} = x_{ij}, b_i, \beta_j)$ denote the models for the selection and rating procedure respectively. Note that in the previous formulation we allow the selection procedure to depend on the parameters that characterize the rating process ($\beta_j$) and also on other selection-specific parameters ($\alpha_j$). It is easy to see that, under conditional independence, (4.6) takes the simpler form

$$P(Y_j = 1|\beta_j) = \int P(Y_{ij} = 1|X_{ij} = 1, b_i, \beta_j)\ \phi(b_i|0, \sigma_b^2)\, db_i. \tag{4.8}$$

Expression (4.8) does not depend on the selection procedure and the estimation of the success probabilities is reduced to the estimation of the clusters effect and the variance component $\sigma_b^2$. However, even though the selection procedure does not explicitly appear in (4.8), one may need to take it into account when estimating the $\beta_j$s and $\sigma_b^2$.

In fact, one estimates these parameters using the complete data $\boldsymbol{Y}_i, \boldsymbol{X}_i \in \{0,1\}^N$. The vector of ratings can be decomposed as $\boldsymbol{Y}_i = (\boldsymbol{Y}_{0i}^T, \boldsymbol{Y}_{1i}^T)^T$, where $\boldsymbol{Y}_{1i} \in \{0,1\}^{N_i}$ is the sub-vector associated with the clusters the expert evaluated, $\boldsymbol{Y}_{0i}^T$ is the obvious complement and $N_i = \mathbf{1}^T \boldsymbol{X}_i$. The joint distribution of $(\boldsymbol{Y}_i, \boldsymbol{X}_i, a_i, b_i)$ takes the form

$$\begin{aligned} &P(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{X}_i = \boldsymbol{x}_i, a_i, b_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\ &\quad = P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i, b_i, \boldsymbol{\beta})\, P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha})\, \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}). \end{aligned}$$

Under the conditional independence assumption,
$P(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i, b_i, \boldsymbol{\beta}) = P(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta})$ and, if one further assumes that conditionally on the $b_i$ the components of the response vector $\boldsymbol{Y}_i$ are independent, then $P(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta}) = P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta})\, P(\boldsymbol{Y}_{0i} = \boldsymbol{y}_{0i}|b_i, \boldsymbol{\beta})$ and

$$\begin{aligned} &P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i, a_i, b_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}) \\ &= \sum_{\boldsymbol{y}_{0i}} P(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\beta})\, P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha})\, \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\ &= \sum_{\boldsymbol{y}_{0i}} P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta})\, P(\boldsymbol{Y}_{0i} = \boldsymbol{y}_{0i}|b_i, \boldsymbol{\beta})\, P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha})\, \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\ &= P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta})\, P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha})\, \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}), \\ &= \left[\prod_j^{N_i} P(Y_{1ij} = y_{1ij}|b_i, \beta_j)\right] \left[\prod_j^{N} P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j)\right] \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}). \end{aligned}$$

Marginally, the previous equations lead to

$$P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i},\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}\right) \tag{4.9}$$
$$=\int\int P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i}|b_i,\boldsymbol{\beta}\right)P\left(\boldsymbol{X}_i=\boldsymbol{x}_i|a_i,\boldsymbol{\beta},\boldsymbol{\alpha}\right)\,\phi(a_i,b_i|\mathbf{0},\boldsymbol{\Sigma})\,da_i db_i,$$

and the likelihood emerging from (4.9) takes the form

$$L\left(\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2\right)=\prod_i^n P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i},\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\beta},\boldsymbol{\alpha},\boldsymbol{\Sigma}\right). \tag{4.10}$$

Using the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\boldsymbol{\alpha}}_n$, $\widehat{\sigma}^2_{bn}$, under conditional independence, one can estimate the probabilities of success by substituting $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\sigma}^2_{bn}$ into (4.8). Note, however, that to estimate $\boldsymbol{\beta}$, $\sigma_b^2$, one may need to explicitly model the selection process. An important special instance where the selection mechanism can be ignored is when the selection and rating processes are also marginally independent, i.e., when $\phi(a_i,b_i|\mathbf{0},\boldsymbol{\Sigma})=\phi(a_i|0,\sigma_a^2)\phi(b_i|0,\sigma_b^2)$ and have a disjoint parametric space. In fact, under these assumptions (4.9) simplifies to

$$P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i},\boldsymbol{X}_i=\boldsymbol{x}_i|\boldsymbol{\beta},\boldsymbol{\alpha},\sigma^2\right)$$
$$=\int P\left(\boldsymbol{X}_i=\boldsymbol{x}_i|a_i,\boldsymbol{\alpha}\right)\,\phi(a_i|0,\sigma_a^2)\,da_i\int P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i}|b_i,\boldsymbol{\beta}\right)\,\phi(b_i|0,\sigma_b^2)\,db_i.$$

Consequently, regarding the parameters of interest $\boldsymbol{\beta}$ and $\sigma_b^2$, the contribution of expert $i$ to the likelihood becomes

$$\int P\left(\boldsymbol{Y}_{1i}=\boldsymbol{y}_{1i}|b_i,\boldsymbol{\beta}\right)\,\phi(b_i|0,\sigma_b^2)\,db_i,=\int\left[\prod_j P\left(Y_{1ij}=y_{1ij}|b_i,\beta_j\right)\right]\,\phi(b_i|0,\sigma_b^2)\,db_i.$$

The previous expression is the contribution of expert $i$ to the likelihood when the selection mechanism has been discarded. Therefore, in this scenario, if conditional independence holds, the selection procedure can be fully ignored.

Importantly, such a scenario will result if a random allocation of the clusters to experts is implemented, where the experts have not influence whatsoever on the selection process. The previous discussion shows that fully random allocation is a powerful tool not only to avoid selection bias, by guaranteeing conditional independence, but also to considerably simplify the analysis by making the selection mechanism ignorable for the estimation of the parameters.

## 4.3 Simulation Study

To numerically evaluate the ignorability of the selection procedure and the impact of selection bias on the assessments, a simulation study was designed. The data

were generated mimicking the main characteristics encountered in the case study. Nonetheless, the size of the simulated data sets were chosen so that model (4.1) could be fitted using maximum likelihood. To that effect, data for only 50 ($N$) clusters and 147 experts ($n$) were generated. The random expert effects $b_i$ were sampled from a $N(0, 10)$. Moreover, the values of the parameters characterizing the cluster effects $\beta_j$ were generated once from a $N(0, 2)$ and then held fixed throughout all simulations. The selection and rating probabilities were computed using the following models

$$\text{logit}\left[P\left(X_{ij} = 1|b_i\right)\right] = b_i, \tag{4.11}$$

$$\text{logit}\left[P(Y_{ij} = 1|b_i)\right] = b_i + \beta_j. \tag{4.12}$$

Notice that models (4.11) and (4.12) are a special case of the general modeling framework introduced in Section 4.2.1. In fact, to simplify the computational burden and improve numerical stability, we considered the situation in which the selection and rating procedures shared a common random effect or, equivalently, $\text{corr}(a_i, b_i) = 1$.

In the previous setting, like in the case study, some experts will tend to evaluate a large number of clusters whereas others will tend to evaluate only a reduced number of them. Note further that the rating process does not depend on the selection procedure, i.e.,

$$P\left(Y_{ij} = 1|X_{ij} = 1, b_i, \beta_j\right) = P\left(Y_{ij} = 1|X_{ij} = 0, b_i, \beta_j\right) = P\left(Y_{ij} = 1|b_i\right),$$

and, therefore, there is no selection bias. In total, 200 data sets were generated and analyzed using model (4.1). Subsequently, the success probability of each cluster was estimated by plugging the necessary maximum likelihood estimators into (4.2). The integral was approximated as

$$P_{S0} = P\left(Y_k = 1\right) = \sum_{q=1}^{Q} \frac{\exp\left(\beta_k + b_q\right)}{1 + \exp\left(\beta_k + b_q\right)},$$

where $Q = 10,000$ and $b_q \sim N(0, \widehat{\sigma}^2)$ when using the $\widehat{\beta}_k$ values estimated from model (4.1) and $b_q \sim N(0, 10)$ when using the true $\beta_k$ values. Table 4.1 summarizes the main results and the clusters are ordered decreasingly according to their true probability of success. Clearly, ignoring the selection procedure can have a huge impact on the estimators $\widehat{\beta}_k$ and, consequently, on the estimates of the success probabilities. Indeed, using the estimated probability of success $\widehat{P}_{S0}$, cluster 32 would be considered the most promising one whereas, in reality, it should be ranked eighth, taking into account its true probability of success. These findings unequivocally showed that ignoring the selection process, when estimating the model parameters and the

probabilities of success, may be extremely misleading even in the absence of selection bias.

Further, we studied a scenario in which selection bias was present. To that effect we considered the following rating mechanism

$$\text{logit}\left[P(Y_{ik}=1|X_{ik}=x_{ik},b_i)\right]=\begin{cases}\beta_k+b_i & \text{if } x_{ik}=1,\\ \beta_k+b_i-0.223 & \text{if } x_{ik}=0.\end{cases} \tag{4.13}$$

Essentially, (4.13) implies that, for every expert $i$, the odds of rating a cluster as 1 is 25% larger when the cluster is evaluated than when it is not. The values of the true success probabilities in this scenario, computed as (4.7), are given under the column $P_{S1}$ in Table 4.1. Note that, even if one can avoid bias when estimating $\beta_k$ and $\sigma^2$, a comparison between $P_{S0}$ and $P_{S1}$ clearly shows that, in the presence of selection bias, a naive use of (4.2) would lead to an overestimation of the true probabilities of success, as it was stated in Section 4.2.1.

In a second simulation study, we took into account the selection process when estimating the parameters of interest. Basically, we maximized likelihood (4.10) considering the selection model $\text{logit}\left[P\left(X_{ij}=1|b_i\right)\right]=b_i+\alpha$. The setting was essentially the same as before but to alleviate the computational burden only 10 clusters were now considered. The results are presented in Table 4.2. Once more, the naive approach that ignores the selection process led to biased estimates for the cluster effects, the variance component and the probabilities of success. Importantly, for some clusters, the relative bias in the estimated probability of success was as large as 25%. Further, when the selection procedure was incorporated into likelihood (4.10), the bias disappeared and the probabilities of success were always accurately estimated. Additional simulations (not shown) with a reduced number of 50 experts confirmed these conclusions.

## 4.4 Case Study Revisited

The case study introduced in Section 2.1 was reanalyzed taking into account the selection process by maximizing likelihood (4.10) with $\text{logit}\left[P\left(X_{ij}=1|a_i\right)\right]=\alpha_j+a_i$. The integrals in (4.10) were estimated with the less efficient Laplace instead of the more efficient adaptive quadrature technique since the complexity of the model led to convergence problems when the latter was used. Though it was possible to fit a less complex model where $\text{Corr}(b_i,a_1)=1$ using adaptive quadrature, Figure 4.1 shows that the assumption implied by this model i.e., experts who rate more clusters

Table 4.1: *Simulation results. ID: cluster id; $\beta_k$: true cluster effect on the rating process, $P_{S0}$ and $P_{S1}$ are success probabilities based on $\beta_k$ (i.e., no model is fitted to estimate $\beta_k$):$P_{S0}$ accounts for the contribution of clusters that were not rated by assuming that they have low success probability than the rated ones while $P_{S1}$ ignores contribution of clusters that were not rated. $\widehat{\beta}_k$ estimates $\beta_k$ using the naive model (model that ignores the selection process) when conditional on the expert, the selection and rating processes are independent and $\widehat{P}_{S0}$ are the corresponding probabilities.*

| | True values | | | Naive | |
|---|---|---|---|---|---|
| ID | $\beta_k$ | $P_{S1}$ | $P_{S0}$ | $\widehat{\beta}_k$ | $\widehat{P}_{S0}$ |
| 3 | 4.326 | 0.858 | 0.865 | 2.338 | 0.746 |
| 1 | 3.602 | 0.813 | 0.821 | -0.259 | 0.471 |
| 33 | 3.518 | 0.807 | 0.815 | -0.320 | 0.463 |
| 47 | 3.434 | 0.801 | 0.809 | 3.146 | 0.808 |
| 50 | 3.037 | 0.772 | 0.781 | 1.683 | 0.684 |
| 27 | 2.127 | 0.696 | 0.706 | -1.216 | 0.364 |
| 30 | 2.059 | 0.690 | 0.700 | 1.272 | 0.642 |
| 32 | 2.056 | 0.690 | 0.700 | 10.228 | 0.947 |
| 14 | 1.892 | 0.675 | 0.685 | 2.374 | 0.749 |
| 7 | 1.701 | 0.657 | 0.668 | 1.591 | 0.676 |
| 9 | 1.505 | 0.639 | 0.650 | 3.366 | 0.804 |
| 48 | 1.369 | 0.625 | 0.637 | 1.950 | 0.711 |
| 10 | 1.293 | 0.618 | 0.629 | 2.581 | 0.767 |
| 21 | 1.032 | 0.592 | 0.604 | 1.690 | 0.685 |
| 11 | 0.876 | 0.577 | 0.588 | -1.637 | 0.320 |
| 26 | 0.873 | 0.577 | 0.588 | 4.348 | 0.863 |
| 15 | 0.685 | 0.558 | 0.569 | 1.671 | 0.683 |
| 13 | 0.602 | 0.549 | 0.561 | 4.249 | 0.851 |
| 4 | 0.582 | 0.547 | 0.559 | 1.827 | 0.698 |
| 42 | 0.389 | 0.528 | 0.540 | 1.314 | 0.646 |
| $\sigma$ | 10.00 | | | 9.080 | |

Table 4.2: *Simulation results.* $\beta_k$*: true value used to generate the data;* $P_{S0}$*: true probability of success. The mean of the estimated values are denoted using the hat symbol. The estimates are obtained using the naive approach that ignores the selection process and the joint model that takes this process into account.*

| | True values | | Naive | | Joint Model | |
|---|---|---|---|---|---|---|
| cid | $\beta_k$ | $P_{S0}$ | $\widehat{\beta}_k$ | $\widehat{P}_{S0}$ | $\widehat{\beta}_k$ | $\widehat{P}_{S0}$ |
| 1 | 3.60 | 0.84 | 6.16 | 0.95 | 5.02 | 0.85 |
| 2 | -1.98 | 0.29 | -0.96 | 0.37 | -2.01 | 0.29 |
| 3 | 4.33 | 0.88 | 9.58 | 0.97 | 7.96 | 0.90 |
| 4 | 0.58 | 0.56 | 1.57 | 0.70 | 0.59 | 0.56 |
| 5 | 0.11 | 0.51 | 1.07 | 0.64 | 0.10 | 0.51 |
| 6 | -0.53 | 0.44 | 0.45 | 0.56 | -0.54 | 0.44 |
| 7 | 1.70 | 0.68 | 2.75 | 0.82 | 1.73 | 0.68 |
| 8 | -0.10 | 0.49 | 0.89 | 0.62 | -0.08 | 0.49 |
| 9 | 1.51 | 0.66 | 2.51 | 0.80 | 1.51 | 0.66 |
| 10 | 1.29 | 0.64 | 2.29 | 0.78 | 1.31 | 0.64 |
| $\widehat{\sigma}^2$ | 10.00 | | 7.103 | | 10.28 | |

tend to give more positive ratings, is not viable. We preferred to estimate the more plausible model rather than precisely estimating the wrong model. The main results are presented in the second part of Table 4.3 (under the 'Joint Model' columns). The 'Naive' approach assumes absence of selection bias and thus fully ignores the selection process while the 'Joint Model' approach assumes conditional absence of selection bias and follows a different path to estimate the parameters of interest, that takes the selection process into account.

Notably the variance estimate is very high ($\approx$ 20), a possible consequence of using Laplace estimation. The estimated probabilities from the two methods are clearly different. In general the joint model produces lower estimates, for example, the success probability for cluster ranked third is 0.87 and 0.72 for naive and joint models respectively. Unlike with the naive approach, the results obtained from the joint model seem to reflect a rather skeptical opinion regarding the potential of the clusters. These results are not comparable to those in Chapter 3 since different estimation techniques were used.

Table 4.4 gives results from joint models with the selection process modeled as, $\text{logit}\left[P\left(X_{ij}=1|a_i\right)\right] = a_i + \beta_j$ and $\text{logit}\left[P\left(X_{ij}=1|a_i\right)\right] = a_i + \alpha$. Results from the model where $\text{logit}\left[P\left(X_{ij}=1|a_i\right)\right] = a_i + \beta_j$ are of different magnitude from all the other models (e.g. success probability for second ranked cluster is 0.29). This selection model is indeed restrictive as it implies positive correlation between $X$ and

Table 4.3: *Estimates for the fixed effects and probabilities of success of the top* 20 *clusters ranked according to the Naive model (analysis that ignores selection bias) and results from the joint model that accounts for selection process.* $\hat{\beta}$ *and* $R$ *are the estimated cluster-effect and ranks respectively.* $\widehat{P}$ *is the estimated probability of success, while lcl and ucl are the corresponding lower and upper 95% confidence limits.*

| | Naive | | | | | Joint | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ID | $R$ | $\hat{\beta}$ | $\widehat{\text{P}}$ | lcl | ucl | $R$ | $\hat{\beta}$ | $\widehat{\text{P}}$ | lcl | ucl |
| 265222 | 1 | 2.52 | 0.94 | 0.78 | 0.98 | 3 | 2.67 | 0.72 | 0.45 | 0.89 |
| 295061 | 2 | 3.83 | 0.92 | 0.66 | 0.98 | 4 | 2.61 | 0.71 | 0.48 | 0.87 |
| 359957 | 3 | 0.49 | 0.87 | 0.72 | 0.94 | 330 | -0.25 | 0.48 | 0.18 | 0.79 |
| 69850 | 4 | 1.07 | 0.82 | 0.33 | 0.97 | 182 | 0.11 | 0.50 | 0.23 | 0.77 |
| 84163 | 5 | 5.24 | 0.77 | 0.41 | 0.97 | 9 | 1.83 | 0.65 | 0.21 | 0.97 |
| 296443 | 6 | 2.59 | 0.76 | 0.49 | 0.93 | 10 | 1.62 | 0.64 | 0.33 | 0.87 |
| 7451 | 7 | 1.28 | 0.74 | 0.16 | 0.96 | 55 | 0.66 | 0.56 | 0.24 | 0.81 |
| 277619 | 8 | 1.65 | 0.73 | 0.41 | 0.94 | 89 | 0.44 | 0.54 | 0.17 | 0.87 |
| 315928 | 9 | 2.04 | 0.72 | 0.37 | 0.92 | 14 | 1.47 | 0.62 | 0.28 | 0.83 |
| 296535 | 10 | 2.77 | 0.71 | 0.48 | 0.87 | 5 | 2.37 | 0.70 | 0.38 | 0.91 |
| 313914 | 11 | 2.18 | 0.70 | 0.40 | 0.89 | 7 | 2.06 | 0.68 | 0.28 | 0.91 |
| 277774 | 12 | 2.20 | 0.69 | 0.43 | 0.87 | 20 | 1.30 | 0.61 | 0.37 | 0.81 |
| 178994 | 13 | 1.85 | 0.68 | 0.45 | 0.84 | 11 | 1.57 | 0.64 | 0.34 | 0.84 |
| 296560 | 14 | 1.89 | 0.66 | 0.43 | 0.83 | 8 | 1.86 | 0.66 | 0.39 | 0.85 |
| 464822 | 15 | 1.21 | 0.66 | 0.43 | 0.83 | 72 | 0.56 | 0.55 | 0.31 | 0.77 |
| 265441 | 16 | 1.87 | 0.65 | 0.41 | 0.86 | 15 | 1.44 | 0.62 | 0.34 | 0.85 |
| 292805 | 17 | 1.47 | 0.65 | 0.38 | 0.84 | 19 | 1.20 | 0.61 | 0.29 | 0.84 |
| 432169 | 18 | 1.45 | 0.64 | 0.35 | 0.86 | 1 | 6.26 | 0.91 | 0.51 | 0.99 |
| 292579 | 19 | 1.85 | 0.64 | 0.24 | 0.90 | 13 | 1.50 | 0.63 | 0.21 | 0.89 |
| 278927 | 20 | 1.30 | 0.63 | 0.41 | 0.81 | 76 | 0.51 | 0.54 | 0.31 | 0.76 |
| $\widehat{\sigma}^2$ | | 20.02 | | | | | 18.61 | | | |

$Y$ not only across the raters but also across the clusters.

When the joint model assuming $\text{Corr}(a_i, b_i) = 1$ is used, the differences between the two methods are more striking. The joint model approach produces much lower estimates of the success probabilities and leads to a completely different ranking of the clusters. Additionally, it also produces a much smaller estimate of the between-experts variability. This trend was also evident with other selection models, for instance, when the selection process was modeled as $\text{logit}\left[P\left(X_{ij} = 1|b_i\right)\right] = b_i + \beta_j$, even smaller estimates were obtained for the success probabilities (results available in Appendix A. These and other additional models open the possibility of a carefully done sensitivity analysis. Such is necessary because the presence of untestable assumption cannot be

Table 4.4: *Results for the top 20 clusters ranked using the naive approach that ignores the selection process. Given are: estimated cluster effect ($\widehat{\beta}_k$); rank assigned to the cluster according to its probability of success obtained by fitting joint models whose selection models are given in the first row.*

| | $\text{logit}\left[P(X_{ij}=1\|a_i)\right]=\beta_j+a_i$ | | | $\text{logit}\left[P(X_{ij}=1\|a_i)\right]=\alpha+a_i$ | | |
|---|---|---|---|---|---|---|
| ID | $\widehat{\beta}_k$ | Rank | Estimated Probability | $\widehat{\beta}_k$ | Rank | Estimated Probability |
| 265222 | -2.51 | 186 | 0.25 | 2.68 | 2 | 0.72 |
| 295061 | -2.68 | 721 | 0.23 | 3.20 | 1 | 0.76 |
| 359957 | -2.95 | 4386 | 0.21 | -0.03 | 172 | 0.49 |
| 69850 | -2.75 | 1259 | 0.23 | 0.65 | 36 | 0.56 |
| 84163 | -3.75 | 20152 | 0.15 | 1.84 | 8 | 0.65 |
| 296443 | -3.31 | 12333 | 0.18 | 1.62 | 10 | 0.64 |
| 7451 | -2.92 | 3759 | 0.21 | 0.50 | 61 | 0.54 |
| 277619 | -2.91 | 3180 | 0.21 | 0.92 | 21 | 0.58 |
| 315928 | -2.47 | 172 | 0.25 | 1.97 | 4 | 0.67 |
| 296535 | -2.91 | 3542 | 0.21 | 2.33 | 3 | 0.70 |
| 313914 | -3.61 | 18883 | 0.16 | 1.94 | 5 | 0.66 |
| 277774 | -2.78 | 1291 | 0.23 | 1.86 | 7 | 0.66 |
| 178994 | -2.05 | 2 | 0.29 | 1.45 | 13 | 0.63 |
| 296560 | -2.56 | 280 | 0.24 | 1.84 | 9 | 0.65 |
| 464822 | -2.89 | 2840 | 0.21 | 0.76 | 34 | 0.56 |
| 265441 | -3.36 | 13757 | 0.18 | 1.53 | 12 | 0.63 |
| 292805 | -2.66 | 723 | 0.23 | 1.30 | 16 | 0.61 |
| 432169 | -3.16 | 9361 | 0.19 | 1.38 | 15 | 0.61 |
| 292579 | -3.06 | 6162 | 0.20 | 1.85 | 6 | 0.66 |
| 278927 | -2.61 | 430 | 0.24 | 0.78 | 31 | 0.57 |
| $\widehat{\sigma}^2$ | 10.11 | | | 18.87 | | |

avoided.

The findings presented in Section 4.2 indicate that the joint-model approach gives a more reliable picture of reality and, Figure 4.1 suggest that the simple model where $\text{Corr}(a_i, b_i) = 1$ is not appropriate, thus we are inclined to put more weight on the results from the joint model in Table 4.3. Assessment of fit of the models is a little more difficult than usual, because we have a non-likelihood method. One could assess the fit in subsets, but then the question about overall fit would still remain. Proper methodology for model fit with pseudo-likelihood methods requires further research. Furthermore, the focus here is on sensitivity with respect to untestable assumptions.

Clearly, a very careful discussion incorporating domain-specific knowledge, will be needed before a final conclusions can be drawn from this study.

## 4.5 Discussion

The topic studied here can be related to other statistical fields and perhaps the most evident connection is with missing data analysis. Indeed, like many problems from areas like hierarchical models (Lindstrom and Bates, 1988), causal inference, and treatment compliance (Holland, 1986), selection bias could also be framed within a missing data context. To illustrate this connection using a simpler notation, let us focus on the special case in which the selection and rating procedures shared a common random effect. Conditioning on the expert effect, one could think of the selection and rating procedures introduced in Section 4.2, as analogous to the pattern mixture framework often use to handle missing observations (Molenberghs and Kenward, 2007). Similarly, the condition used to define selection bias in Section 4.2 is closely related to the concept of *missing not at random* (MNAR) that appears in the classical missing data taxonomy (Rubin, 1976; Kenward and Carpenter, 2007; Molenberghs and Kenward, 2007), and which means that the missing-data mechanism is related to unobserved outcomes, in addition to observed outcomes and covariates. To exemplify this, consider the expression

$$P\left(Y_{ij} = y_{ij} | X_{ij} = x_{ij}, b_i\right) = P\left(X_{ij} = x_{ij} | Y_{ij} = y_{ij}, b_i\right) \frac{P\left(Y_{ij} = y_{ij} | b_i\right)}{P\left(X_{ij} = x_{ij} | b_i\right)}. \quad (4.14)$$

If the probability of not evaluating a cluster is independent of its (unobserved) rating, then we have $P\left(X_{ij} = 0 | Y_{ij} = y_{ij}, b_i\right) = P\left(X_{ij} = 0 | b_i\right)$, which is the definition of the *Missing At Random* mechanism (MAR) in the Rubin taxonomy (Rubin, 1976). MAR means that, given observed outcomes and covariates, missingness does not further depend on unobserved ones. It is easy to see that (4.14) and the subsequent

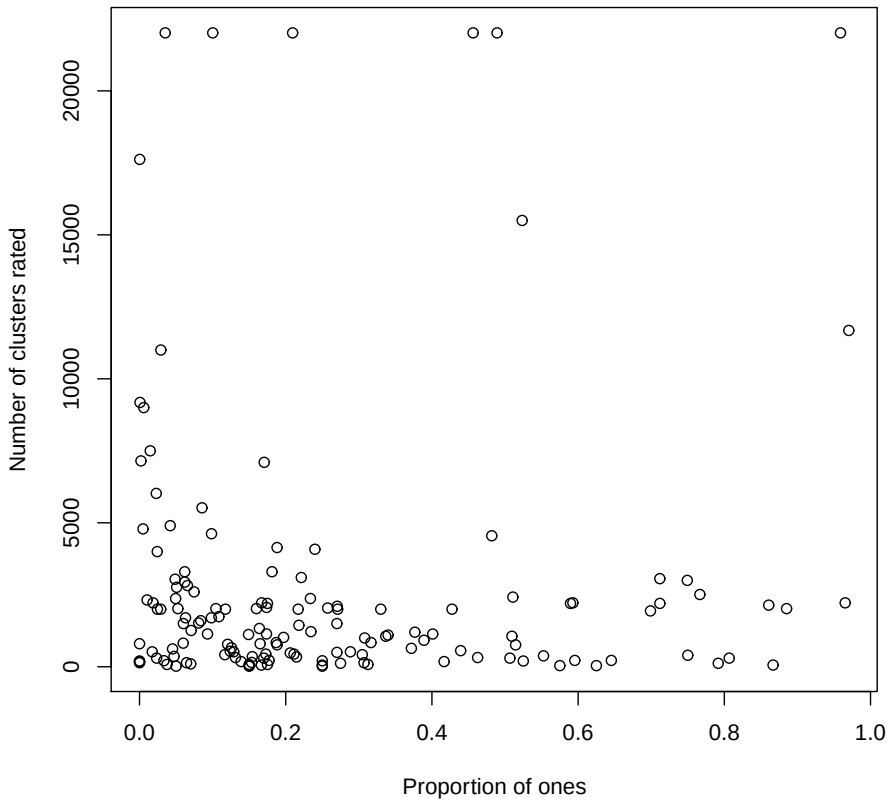


Figure 4.1: *A scatter plot for the number of clusters rated and the proportion of clusters recommeded by each expert*

expressions imply

$$P\left(Y_{ij} = y_{ij} | X_{ij} = 1, b_i\right) = P\left(Y_{ij} = y_{ij} | X_{ij} = 0, b_i\right) = P\left(Y_{ij} = y_{ij} | b_i\right),$$

and, therefore, the absence of selection bias can be seen as an MAR process, given the expert. Moreover, the conditional independence assumption for the rating and selection procedure introduced in Section 4.2, is closely related to the shared parameter modeling (SPM) framework, regularly used to describe a MNAR mechanism (Follmann and Wu, 1995; Little, 1995). This relationship with the SPM explains why, unlike under MAR, where the likelihood paradigm implies ignorability, in the context studied in this manuscript even in absence of selection bias the selection procedure will

often be non-ignorable. The reason for this important difference is that the selection and rating procedures share a common random effect $b_i$ and, therefore, marginally independence does not hold. In the most general case when the selection and rating processes are governed by two correlated random effects $a_i$ and $b_i$ parallels can be drawn with the so-called generalized SPM (Creemers *et al.*, 2011), but this will not be explored further here.

It has been shown that in a missing data problem the data at hand do not provide enough information to discriminate between MAR and MNAR (Molenberghs *et al.*, 2008). Likewise, the data at hand will not provide enough information to discard the presence of selection bias if the assignment mechanism was non-random or had the potential to be influenced by the experts. One could, however, conceive a sensitivity analysis to evaluate the robustness of the conclusions with respect to the potential presence of selection bias.

# Chapter 5

# A New modeling Approach for Quantifying Expert Opinion in the Drug Discovery Process

In Chapter 4 we studied the conditions under which the selection process (process by which experts are assigned the clusters to rate) could be safely ignored when analyzing the data in Section 2.1 and showed that inappropriately ignoring it may seriously threaten the validity of the study. As a consequence, one often needs to jointly model the rating and selection processes in order to avoid bias. Ideally, one would like to know all the factors influencing the selection process beforehand. However, in practice, such information is seldom available and making assumptions on the selection process is almost inescapable, and if these assumptions are wrong, estimates and inferences may be wrong as well. In this chapter, a new modeling technique that produces valid estimates even under misspecification of the assumptions made on the selection process is introduced. This is unlike the joint model introduced in Chapter 4, which is sensitive to misspecification of the assumptions.

## 5.1 Introduction

We shall consider two approaches to account for the selection process. In the first approach, two generalized linear mixed models (GLMM) are used to describe the rating and selection processes and it is assumed that, given some random effect common to both models, both processes are independent. We shall refer to this approach as the joint modeling approach. The previous conditional independence assumption is at the core of the so-called shared-parameter models commonly used, for instance, to jointly model longitudinal and survival outcomes or in the analysis of missing not at random data (Rizopoulos, 2012; Vonesh, Green, and Schluchter, 2006; Rizopoulos, Verbeke, and Molenberghs, 2008; Follmann and Wu, 1995). In the present work the aforementioned assumption simplifies the joint distribution of the rating and selection processes, facilitating the joint fit of both models.

The previous approach hinges on the assumption that the distribution for the selection process is correctly specified. In general, if the selection model is misspecified then the estimates of the parameters in the rating model may be biased and inferential procedures, like confidence intervals, may be affected as well. Therefore, a sensitivity analysis to assess the stability of the results is always highly recommended (Geneletti *et al.*, 2011).

Our second approach is based on the so-called combined model introduced by Booth *et al.* (2003) and Molenberghs *et al.* (2010) for members of the exponential family, where an extra set of random effects is used to account for overdispersion in correlated outcomes. Similarly, in this work, we propose to take into account the selection process by adding a new set of random effects to the rating model. It is important to point out that, although the combined model has been shown to improve model fit in overdispersed correlated data, its usefulness to account for selection bias is yet to be investigated.

We extensively study the performance of both approaches via simulation. Our results show that the combined model could be a robust alternative to the joint model when analyzing the data. However, as one would expect, if the selection model is correctly specified then the joint model will deliver better results. Therefore, we think that the combined model may serve two purposes: (i) it may be a reliable tool for sensitivity analysis and (ii) when there are doubts regarding the validity of the selection model, it may be a safe alternative on which to base inferences.

## 5.2 The Joint Modeling Approach

Recall that the marginal probability of success for cluster $C_j$ was obtained as,

$$P\left(Y_j = 1|\beta_j\right) = \int P\left(Y_{ij} = 1|b_i, \beta_j\right)\, \phi(b_i|0, \sigma_b^2)\, db_i, \tag{5.1}$$

where $\phi(b_i|0, \sigma_b^2)$ denotes a normal density with mean zero and variance $\sigma_b^2$. Even though the selection procedure does not explicitly appear in (5.1), one may still need to take it into account when estimating the $\beta_j$s and $\sigma^2$ in order to avoid bias. In fact, one estimates these parameters using the complete data $\boldsymbol{Y}_i, \boldsymbol{X}_i \in \{0,1\}^N$. The vector of ratings can be decomposed as $\boldsymbol{Y}_i = (\boldsymbol{Y}_{0i}^T, \boldsymbol{Y}_{1i}^T)^T$, where $\boldsymbol{Y}_{1i} \in \{0,1\}^{N_i}$ is the sub-vector associated with the clusters the expert evaluated, $\boldsymbol{Y}_{0i}^T$ is the obvious complement and $N_i = \mathbf{1}^T \boldsymbol{X}_i$. The joint distribution of $(\boldsymbol{Y}_i, \boldsymbol{X}_i, a_i, b_i)$ takes the form

$$\begin{aligned} &P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{X}_i = \boldsymbol{x}_i, a_i, b_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) \\ &\quad = P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|\boldsymbol{X}_i = \boldsymbol{x}_i, b_i, \boldsymbol{\beta}\right) P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) \phi(a_i, b_i|\mathbf{0}, \boldsymbol{\Sigma}). \end{aligned} \tag{5.2}$$

with

$$P(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}|b_i, \boldsymbol{\beta}) = \prod_j^{N_i} P(Y_{1ij} = y_{1ij}|b_i, \beta_j)$$

and, similarly,

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \prod_j^{N} P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j),$$

as shown in Chapter 4. In the previous formulation the selection procedure given by $P(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j)$ is allowed to depend on the parameters that characterize the rating process $(\beta_j)$ and also on other selection-specific parameters $(\alpha_j)$. The likelihood emerging from (5.2) is

$$L\left(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right) = \prod_i^n P\left(\boldsymbol{Y}_{1i} = \boldsymbol{y}_{1i}, \boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\Sigma}\right). \tag{5.3}$$

Using the maximum likelihood estimators $\widehat{\boldsymbol{\beta}}_n$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\Sigma}}$, one can estimate the probabilities of success by substituting $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_b^2$ into (5.1). Note, however, that to estimate $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, one may need to explicitly model the selection process. In Chapter 4, we showed that if the selection probability is independent of the rater specific effect $(b_i, a_i)$ and the rating parameters $\boldsymbol{\beta}$, i.e., $P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|a_i, \boldsymbol{\beta}, \boldsymbol{\alpha}\right) = P\left(\boldsymbol{X}_i = \boldsymbol{x}_i|\boldsymbol{\alpha}\right)$ then the selection mechanism can be safely ignored. This setting will result, for instance, if a fully random allocation of the clusters to raters is implemented, so that the raters have no

influence on the selection process, else a selection model will need to be incorporated into the analysis in order to avoid bias and this raises questions about the impact of misspecifying this model on the estimates of the parameters of interest.

## 5.3 Combined Model Approach

The combined model follows a different path for estimating the parameters of interest, namely, the $\beta_j$s and $\sigma^2$. To introduce this approach let us first notice that, in the joint model, the selection process $P\left(X_{ij} = x_{ij}|a_i, \beta_j, \alpha_j\right)$ appearing in the integrand in (5.2) is directly modeled using, for instance, a GLMM. Alternatively, we now propose to account for the extra variability emanating from the selection process by introducing a new set of random effects $\theta_{ij}$. Essentially, we propose to work with the conditional distribution

$$\begin{aligned} f\left(\boldsymbol{Y}_i = \boldsymbol{y}_i, \boldsymbol{\theta}_i|b_i\right) =& P\left(\boldsymbol{Y}_i = \boldsymbol{y}_i|b_i, \boldsymbol{\theta}_i\right) f\left(\boldsymbol{\theta}_i|b_i\right) \\ =& \prod_j^N P\left(Y_{ij} = y_{ij}|b_i, \theta_{ij}\right) f\left(\theta_{ij}|b_i\right), \end{aligned} \tag{5.4}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iN})^T$. The previous expression assumes that, conditional on the random effects, the ratings are independent and so are components of $\boldsymbol{\theta}_i$. Since the two sets of random effects are meant to explain different sources of variability, $\boldsymbol{\theta}_i$ and $b_i$ are also assumed to be independent, hence, $f\left(\boldsymbol{\theta}_i|b_i\right) = f\left(\boldsymbol{\theta}_i\right)$. Finally, $\theta_{ij} \sim \text{Beta}\left(\lambda, \tau\right)$ and $Y_{ij}|\theta_{ij}, b_i \sim \text{Bernoulli}(\theta_{ij}\pi_{ij})$ with

$$\pi_{ij} = \frac{\exp(\beta_j + b_i)}{1 + \exp(\beta_j + b_i)}.$$

The previous model directly corresponds to the model introduced by Molenberghs *et al.* (2010). Although there are obvious similarities between the distribution given in (5.4) and the integrand used in expression (5.2), both approaches are fundamentally different. Essentially, the strength of the combined model approach lies in using two sets of random effects, one of which is conjugate to the distribution of the ratings. The conjugate random effects account for the selection process, whereas the normal random effect accounts for the correlation within the set of ratings of a given rater. Often the selection process is not of particular scientific interest and does not need to be exhaustively modeled. Therefore, using random effects to account for it is both desirable and appealing. Considering the previously introduced partition $\boldsymbol{Y}_i = (\boldsymbol{Y}_{0i}^T, \boldsymbol{Y}_{1i}^T)^T$ and the corresponding counterpart $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{0i}^T, \boldsymbol{\theta}_{1i}^T)^T$, expression (5.4)

takes the form

$$f\left(\boldsymbol{Y}_{0i}, \boldsymbol{Y}_{1i}, \boldsymbol{\theta}_{0i}, \boldsymbol{\theta}_{1i} | b_i\right) = P\left(\boldsymbol{Y}_{0i} | \boldsymbol{\theta}_{0i}, b_i\right) P\left(\boldsymbol{Y}_{1i} | \boldsymbol{\theta}_{1i}, b_i\right) f\left(\boldsymbol{\theta}_{0i}, \boldsymbol{\theta}_{1i}\right),$$

and after marginalizing out the subvectors $\boldsymbol{Y}_{0i}$, $\boldsymbol{\theta}_{0i}$ one gets

$$f\left(\boldsymbol{Y}_{1i}, \boldsymbol{\theta}_{1i} | b_i\right) = P\left(\boldsymbol{Y}_{1i} | \boldsymbol{\theta}_{1i}, b_i\right) f\left(\boldsymbol{\theta}_{1i}\right).$$

The parameter estimates are derived using the marginal likelihood obtained after integrating out the random effects $b_i$ and $\boldsymbol{\theta}_{1i}$. This process is carried out in two steps. First after analytically integrating over $\boldsymbol{\theta}_{1i}$ the conditional likelihood contribution for each rater follows as

$$\begin{aligned} L_c^*(\boldsymbol{\beta}, \lambda, \tau, b_i) &= \int f\left(\boldsymbol{Y}_{1i}, \boldsymbol{\theta}_{1i} | b_i\right) \; f\left(\boldsymbol{\theta}_{1i}\right) \, d\boldsymbol{\theta}_{1i}, \\ &= \prod_{j=1}^{N_i} \left\{ \frac{1}{\lambda + \tau} (\pi_{ij}\lambda)^{y_{ij}} \left[(1 - \pi_{ij})\lambda + \tau\right]^{1 - y_{ij}} \right\}, \end{aligned} \tag{5.5}$$

and, eventually, in the second step the marginal likelihood can be obtained by numerically integrating over the normal random effect $b_i$, using readily available statistical software, i.e., the parameter estimates follow from maximizing

$$L_m(\boldsymbol{\beta}, \lambda, \tau, \sigma^2) = \prod_i^n \int L_c^*(\boldsymbol{\beta}, \lambda, \tau, b_i) \phi(b_i | 0, \sigma^2) \, db_i. \tag{5.6}$$

The parameters $\lambda$ and $\tau$ are not simultaneously estimable and to ensure identifiability of the model one parameter has to be fixed. To avoid subjectivity we estimate the ratio $\frac{\lambda}{\tau}$ and not the individual parameters.

## 5.4 Simulation Study

When working with hierarchical models one often has to deal with likelihood functions that do not have a closed form. For instance, combining normal random effects and binary outcomes with logit links leads to an unclosed form for the marginal likelihood and, therefore, one needs to resort to numerical algorithms to compute the maximum likelihood estimators (MLE). Consequently, studying the properties of the MLE theoretically is extremely difficult in many settings and simulation studies become an indispensable tool to compare alternative approaches in these scenarios.

In this work, the data were generated mimicking the case study introduced in Section 2.1, but the size of the simulated data sets were chosen so that both, the

joint and combined models, could be fitted using maximum likelihood. To that effect 147 experts and 15 clusters were considered for the simulations. The fixed-effects $\beta_j$, $\alpha_j$ were sampled once from a $N(0, 2)$ and $N(0, 1)$ respectively and then held fixed throughout all simulations, whereas the random rater specific effects $b_i$ were sampled from $N(0, 10)$. To mimic the random allocation used in the case study a number of clusters $N_i$ was randomly assigned to each rater, with $N_i$ coming from a Poisson(25) and $2 \leq N_i \leq 15$. Finally, the actual clusters evaluated by each rater were defined using the selection process $X_{ij}|b_i \sim$ Bernoulli $(\rho_{ij})$ with $\text{logit}(\rho_{ij}) = \alpha_j + 0.5 \cdot b_i$ and the corresponding ratings $Y_{ij}|b_i$ were generated from a Bernoulli$(\pi_{ij})$ with

$$\pi_{ij} = \frac{\exp(\beta_j + b_i)}{1 + \exp(\beta_j + b_i)} .$$

Using the previous settings, a total of 200 data sets were generated. Three analyses were carried out for each data set and the main results are summarized in Tables 5.1–5.4. In these tables, the column *True* gives always the true value of the corresponding parameter, the column *Combined* refers to results obtained from the combined model introduced in Section 5.3, the column $J(\cdot)$ displays the results obtained from fitting the joint model using the selection probability derived from the logit in brackets and, finally, the column *Naive* presents the results obtained from fitting model without accounting for the different selection probabilities.

The model $j(\alpha_j + 0.5 \cdot b_i)$, which assumes that selection probability of each cluster vary within each expert, and the parameters governing the rating and selection processes are different, is the correctly specified model. In contrast, $j(\beta_j + 0.5 \cdot b_i)$ also postulates different selection probabilities for the clusters but now the parametric space of the rating and selection processes are assumed to be equal. The last model $j(\alpha + 0.5 \cdot b_i)$ presupposes equal selection probabilities for all the clusters. Obviously joint models with different random effects for the rating and selection processes would have been more enlightening, but for computation convenience we used the joint model with the shared random effects.

Tables 5.1–5.2 show that, as expected, when the selection model is correctly specified $(j(\alpha_j + 0.5 \cdot b_i))$ the joint modeling approach produces points estimates of the parameters of interest that are very close to their true values. Nonetheless, when the selection model is misspecified relative biases larger than 200% may appear. Similarly, ignoring the selection process can also be problematic. Indeed, as the results from the naive analysis show, relative biases larger than 400% can be obtained when the selection process is incorrectly ignored. Unlike in the previous cases, the combined model always led to unbiased estimates of the parameters. However, from all the

models considered, it produced the largest standard errors as well. In fact, the misspecified joint models exhibited in some settings large bias and high precision, while the combined model had smaller bias and lower precision.

It is important to point out that highly precise but incorrect estimates could lead to seriously misleading inferences. In fact, as shown in Table 5.3, the fixed effects parameters were estimated with high precision when model $j(\beta_j + 0.5b_i)$ was used, however, the confidence interval coverage for thirteen of them was below 50% and it was even below 10% for seven of them. Similarly, the naive model also exhibited a poor performance with coverage probabilities sometimes far below the pre-specified 95%. In contrast, the combined model always produced confidence intervals with good coverage.

Finally, Table 5.4 displays the true and estimated probabilities of success for every cluster. Here again the combined and correctly specified joint model $[j(\alpha_j + 0.5 \cdot b_i)]$ led to estimate values that are almost equal to the true probabilities. However, the misspecified and naive models produced biased results with relative biases as large as 40% in some scenarios. The joint model $[j(\alpha + 0.5 \cdot b_i)]$ is performing as good as the correctly specified model possibly because values of $\alpha_j \sim N(0, 1)$ do vary wide enough, an extra simulation study with larger variance would be required to confirm this.

The same trend is observed when a different joint model, say, $[j(\beta_j + 0.5 \cdot b_i)]$ is correctly specified (results not shown). The combined model produces estimates that are close to correctly specified model although its performance in presence of selection bias and other forms of joint models is yet to be assessed.

## 5.5 Case Study Analysis

The case study introduced in Section 2.1 was analyzed by using the naive and joint model approaches in Chapter 3 and 4. In this section, the combined model presented in Section 5.3 was also fitted to these data. A summary of the analyses can be found in Table 5.5 where the clusters are ordered according to the results obtained from the naive model. Remarkably, the three approaches lead to strikingly different results. First, notice that the probabilities of success derived from the combined model are relatively smaller than those obtained from the naive and joint methods. Secondly, the ranks given to the clusters by the three approaches also differ in important ways. For instance, the third best cluster according to the naive approach (359,957) received ranks 330 and 88 from the joint and combined models respectively. Moreover, cluster

Table 5.1: *Estimates (standard errors) for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, logit($\rho_{ij}$) = $\alpha_j + 0.5 \cdot b_i$. The column True gives the true values, Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

| $\beta_j$ | True | Combined | J($\alpha_j + 0.5 \cdot b_i$) | J($\beta_j + 0.5 \cdot b_i$) | J($\alpha + 0.5 \cdot b_i$) | Naive |
|---|---|---|---|---|---|---|
| $\beta_1$ | 3.60 | 3.60(0.77) | 3.69 (0.54) | 2.01(0.27) | 3.72 (0.55) | 3.96 (0.57) |
| $\beta_2$ | -1.98 | -1.97(0.66) | -1.98 (0.52) | -1.42(0.22) | -1.95 (0.53) | -1.31 (0.50) |
| $\beta_3$ | 4.33 | 4.33(0.94) | 4.44 (0.59) | 2.49(0.29) | 4.45 (0.60) | 4.70 (0.62) |
| $\beta_4$ | 0.58 | 0.59(0.49) | 0.62 (0.46) | 0.12(0.22) | 0.64 (0.46) | 1.03 (0.45) |
| $\beta_5$ | 0.11 | 0.11(0.49) | 0.15 (0.46) | -0.16(0.22) | 0.14 (0.46) | 0.56 (0.45) |
| $\beta_6$ | -0.53 | -0.52(0.51) | -0.54 (0.46) | -0.53(0.22) | -0.52 (0.46) | -0.02 (0.45) |
| $\beta_7$ | 1.70 | 1.70(0.56) | 1.73 (0.47) | 0.79(0.23) | 1.76 (0.47) | 2.06 (0.47) |
| $\beta_8$ | -0.10 | -0.09(0.50) | -0.14 (0.46) | -0.30(0.22) | -0.12 (0.47) | 0.35 (0.45) |
| $\beta_9$ | 1.51 | 1.52(0.51) | 1.56 (0.46) | 0.71(0.23) | 1.59 (0.46) | 1.90 (0.46) |
| $\beta_{10}$ | 1.29 | 1.30(0.50) | 1.32 (0.46) | 0.56(0.23) | 1.35 (0.47) | 1.71 (0.46) |
| $\beta_{11}$ | 0.88 | 0.91(0.50) | 0.92 (0.46) | 0.32(0.22) | 0.93 (0.46) | 1.28 (0.45) |
| $\beta_{12}$ | -3.52 | -3.49(0.92) | -3.56 (0.64) | -2.27(0.24) | -3.55 (0.64) | -2.64 (0.59) |
| $\beta_{13}$ | 0.60 | 0.61(0.49) | 0.66 (0.46) | 0.12(0.22) | 0.71 (0.46) | 1.08 (0.45) |
| $\beta_{14}$ | 1.89 | 1.90(0.56) | 1.88 (0.47) | 0.91(0.23) | 1.91 (0.47) | 2.24 (0.47) |
| $\beta_{15}$ | 0.68 | 0.71(0.48) | 0.76 (0.46) | 0.22(0.22) | 0.78 (0.46) | 1.17 (0.45) |
| $\sigma^2$ | 10.00 | 10.17(4.31) | 10.52 (2.10) | 6.69(1.19) | 10.53 (2.12) | 8.41 (1.83) |

265,222 ranked first and second by the naive and joint models respectively, was not among the top ten clusters according to the combined model.

Sensitivity of the results with respect to the modeling approach represents a clear dilemma when analyzing this problem. Several strategies could be implemented here, for instance, one could compute the average rank (probability of success) over the different approaches and select those clusters with the largest average rank (probability). On the other hand, given the results of the simulations one could argue that, unlike the naive and joint models, the combined model seems to produce unbiased estimates in most circumstances and, therefore, it should be the core of the decision making process. While we can never be sure that the combined model would fit the data

Table 5.2: *Relative bias for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, logit$(\rho_{ij}) = \alpha_j + 0.5 \cdot b_i$. The column Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

| $\beta_j$ | Combined | J$(\alpha_j + 0.5 \cdot b_i)$ | J$(\beta_j + 0.5 \cdot b_i)$ | J$(\alpha + 0.5 \cdot b_i)$ | Naive |
|---|---|---|---|---|---|
| $\beta_1$ | 0.00 | 0.02 | 0.44 | 0.03 | 0.10 |
| $\beta_2$ | 0.00 | 0.00 | 0.28 | 0.01 | 0.34 |
| $\beta_3$ | 0.00 | 0.03 | 0.42 | 0.03 | 0.09 |
| $\beta_4$ | 0.01 | 0.07 | 0.79 | 0.10 | 0.78 |
| $\beta_5$ | 0.06 | 0.42 | 2.46 | 0.34 | 4.26 |
| $\beta_6$ | 0.01 | 0.03 | 0.01 | 0.02 | 0.95 |
| $\beta_7$ | 0.00 | 0.02 | 0.53 | 0.03 | 0.21 |
| $\beta_8$ | 0.06 | 0.42 | 1.95 | 0.24 | 4.50 |
| $\beta_9$ | 0.01 | 0.03 | 0.53 | 0.06 | 0.26 |
| $\beta_{10}$ | 0.00 | 0.02 | 0.56 | 0.04 | 0.32 |
| $\beta_{11}$ | 0.03 | 0.05 | 0.63 | 0.06 | 0.46 |
| $\beta_{12}$ | 0.01 | 0.01 | 0.36 | 0.01 | 0.25 |
| $\beta_{13}$ | 0.01 | 0.10 | 0.79 | 0.17 | 0.79 |
| $\beta_{14}$ | 0.01 | 0.00 | 0.52 | 0.01 | 0.19 |
| $\beta_{15}$ | 0.03 | 0.10 | 0.68 | 0.14 | 0.71 |
| $\sigma^2$ | 0.02 | 0.05 | 0.33 | 0.05 | 0.16 |

well had the selection process be known, it is useful as a component of a sensitivity analysis. Whatever strategy is finally adopted a careful discussion with the experts in the field would always be advisable in a situation like this one. Eventually, weighting together the quantitative elements emanating from the statistical analysis and more field specific knowledge may help to make an optimal and thoughtful choice.

## 5.6 Discussion

Even in carefully designed studies it is not always possible to avoid bias in the estimates of the parameters of interest. This implies that, when quantifying expert opinion in the drug discovery process, one often needs to jointly fit complex hierar-

chical models describing the selection and rating mechanisms in order to obtain valid estimates. However, in the present work it has been shown that misspecifying the selection model may introduce severe bias in the estimates of the relevant parameters.

We have introduced a new approach using the so-called combined model that accounts for the selection process using a new set of random effects. Simulations results clearly showed that, unlike the naive and joint model approaches, the combined model seems to produce unbiased, although less precise, estimates in most settings. This loss of precision may be seen as the price to pay for the robustness archived by the model.

We believe that even when factors suspected to drive the selection process are known and available, one may still want to use the combined model as a sensitivity tool for the analysis.

Table 5.3: *Confidence interval coverage for the parameters governing the rating process obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, logit*$(\rho_{ij}) = \alpha_j + 0.5 \cdot b_i$*. The column Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.*

| $\beta_j$ | Combined | J($\alpha_j + 0.5 \cdot b_i$) | J($\beta_j + 0.5 \cdot b_i$) | J($\alpha + 0.5 \cdot b_i$) | Naive |
|---|---|---|---|---|---|
| $\beta_1$ | 0.99 | 0.96 | 0.00 | 0.96 | 0.95 |
| $\beta_2$ | 0.95 | 0.95 | 0.32 | 0.96 | 0.76 |
| $\beta_3$ | 0.98 | 0.95 | 0.00 | 0.95 | 0.95 |
| $\beta_4$ | 0.97 | 0.97 | 0.44 | 0.97 | 0.83 |
| $\beta_5$ | 0.97 | 0.96 | 0.74 | 0.97 | 0.85 |
| $\beta_6$ | 0.98 | 0.96 | 0.98 | 0.94 | 0.80 |
| $\beta_7$ | 0.99 | 0.95 | 0.03 | 0.95 | 0.87 |
| $\beta_8$ | 0.97 | 0.94 | 0.82 | 0.94 | 0.81 |
| $\beta_9$ | 0.99 | 0.96 | 0.07 | 0.97 | 0.84 |
| $\beta_{10}$ | 0.99 | 0.97 | 0.10 | 0.97 | 0.91 |
| $\beta_{11}$ | 0.98 | 0.93 | 0.27 | 0.93 | 0.87 |
| $\beta_{12}$ | 0.95 | 0.94 | 0.00 | 0.95 | 0.67 |
| $\beta_{13}$ | 0.96 | 0.96 | 0.45 | 0.95 | 0.87 |
| $\beta_{14}$ | 1.00 | 0.96 | 0.03 | 0.96 | 0.89 |
| $\beta_{15}$ | 0.97 | 0.95 | 0.40 | 0.95 | 0.85 |
| $\sigma^2$ | 0.98 | 0.93 | 0.26 | 0.93 | 0.79 |

Table 5.4: *Estimates for the success probabilities (Relative bias) obtained from the different models fitted to the data. The data were generated by a joint model with selection probability, logit($\rho_{ij}$) = $\alpha_j + 0.5 \cdot b_i$. The column True gives the true values, Combined gives the results obtained from the combined model, 'J' refers to the results obtained from a joint model with the logit of the selection probabilities given in brackets and Naive indicates results from the model that disregards the selection process.Cid identifies the cluster.*

| Rank | Cid. | True | Combined | J($\alpha_j + 0.5 \cdot b_i$) | J($\beta_j + 0.5 \cdot b_i$) | J($\alpha + 0.5 \cdot b_i$) | Naive |
|---|---|---|---|---|---|---|---|
| 1 | 3 | 0.88 | 0.88 (0.00) | 0.89 (0.00) | 0.79 (0.10) | 0.89 (0.00) | 0.91(0.03) |
| 2 | 1 | 0.84 | 0.84 (0.00) | 0.84 (0.00) | 0.74 (0.12) | 0.84 (0.00) | 0.88(0.04) |
| 3 | 14 | 0.70 | 0.70 (0.00) | 0.70 (0.00) | 0.62 (0.12) | 0.70 (0.00) | 0.75(0.07) |
| 4 | 7 | 0.68 | 0.68 (0.00) | 0.68 (0.00) | 0.60 (0.12) | 0.68 (0.00) | 0.73(0.07) |
| 5 | 9 | 0.66 | 0.66 (0.00) | 0.66 (0.00) | 0.59 (0.11) | 0.67 (0.01) | 0.71(0.08) |
| 6 | 10 | 0.64 | 0.64 (0.00) | 0.64 (0.00) | 0.57 (0.11) | 0.64 (0.01) | 0.69(0.08) |
| 7 | 11 | 0.60 | 0.60 (0.01) | 0.60 (0.00) | 0.54 (0.09) | 0.60 (0.01) | 0.65(0.09) |
| 8 | 15 | 0.58 | 0.58 (0.00) | 0.58 (0.01) | 0.53 (0.08) | 0.58 (0.01) | 0.63(0.10) |
| 9 | 13 | 0.57 | 0.57 (0.00) | 0.57 (0.01) | 0.52 (0.09) | 0.57 (0.01) | 0.62(0.10) |
| 10 | 4 | 0.56 | 0.57 (0.00) | 0.57 (0.01) | 0.51 (0.09) | 0.57 (0.01) | 0.62(0.10) |
| 11 | 5 | 0.51 | 0.51 (0.00) | 0.52 (0.01) | 0.48 (0.06) | 0.52 (0.01) | 0.56(0.10) |
| 12 | 8 | 0.49 | 0.49 (0.00) | 0.48 (0.01) | 0.46 (0.06) | 0.49 (0.00) | 0.54(0.11) |
| 13 | 6 | 0.44 | 0.44 (0.01) | 0.44 (0.01) | 0.43 (0.02) | 0.45 (0.01) | 0.50(0.12) |
| 14 | 2 | 0.29 | 0.29 (0.01) | 0.29 (0.01) | 0.33 (0.11) | 0.29 (0.01) | 0.34(0.18) |
| 15 | 12 | 0.17 | 0.17 (0.03) | 0.17 (0.00) | 0.23 (0.41) | 0.17 (0.01) | 0.22(0.32) |

Table 5.5: *Estimated parameters ($\hat{\beta}$), probabilities of success ($\widehat{P}$) and ranks for the top 20 clusters (according to the naive approach) from the case study. The models fitted are: Combined model (Combined), mixed logistic regression (Naive), and joint model with selection probability given by logit$[P(x_{ij} = 1|a_i)] = \alpha_j + a_i$ $[J(\alpha_j + a_i)]$. The column CID gives the cluster id.*

| | Naive | | | $J(\alpha_j + a_i)$ | | | Combined | | |
|---|---|---|---|---|---|---|---|---|---|
| CID | $\hat{\beta}$ | $\widehat{\text{P}}$ | Rank | $\hat{\beta}$ | $\widehat{\text{P}}$ | Rank | $\hat{\beta}$ | $\widehat{\text{P}}$ | Rank |
| 265222 | 2.52 | 0.94 | 1 | 2.67 | 0.72 | 3 | 0.97 | 0.62 | 25 |
| 295061 | 3.83 | 0.92 | 2 | 2.61 | 0.71 | 4 | 1.69 | 0.71 | 1 |
| 359957 | 0.49 | 0.87 | 3 | -0.25 | 0.48 | 330 | 0.71 | 0.59 | 88 |
| 69850 | 1.07 | 0.82 | 4 | 0.11 | 0.50 | 182 | 0.89 | 0.61 | 38 |
| 84163 | 5.24 | 0.77 | 5 | 1.83 | 0.65 | 9 | 1.34 | 0.67 | 6 |
| 296443 | 2.59 | 0.76 | 6 | 1.62 | 0.64 | 10 | 0.55 | 0.57 | 162 |
| 7451 | 1.28 | 0.74 | 7 | 0.66 | 0.56 | 55 | 0.61 | 0.57 | 147 |
| 277619 | 1.65 | 0.73 | 8 | 0.44 | 0.54 | 89 | 0.60 | 0.58 | 138 |
| 315928 | 2.04 | 0.72 | 9 | 1.47 | 0.62 | 14 | 1.26 | 0.66 | 9 |
| 296535 | 2.77 | 0.71 | 10 | 2.37 | 0.70 | 5 | 1.58 | 0.70 | 2 |
| 313914 | 2.18 | 0.70 | 11 | 2.06 | 0.68 | 7 | 1.47 | 0.68 | 4 |
| 277774 | 2.20 | 0.69 | 12 | 1.30 | 0.61 | 20 | 0.98 | 0.62 | 24 |
| 178994 | 1.85 | 0.68 | 13 | 1.57 | 0.64 | 11 | 1.14 | 0.65 | 13 |
| 296560 | 1.89 | 0.66 | 14 | 1.86 | 0.66 | 8 | 1.09 | 0.64 | 15 |
| 464822 | 1.21 | 0.66 | 15 | 0.56 | 0.55 | 72 | 0.90 | 0.61 | 40 |
| 265441 | 1.87 | 0.65 | 16 | 1.44 | 0.62 | 15 | 0.90 | 0.61 | 34 |
| 292805 | 1.47 | 0.65 | 17 | 1.20 | 0.61 | 19 | 1.06 | 0.64 | 20 |
| 432169 | 1.45 | 0.64 | 18 | 6.26 | 0.91 | 1 | 1.01 | 0.63 | 21 |
| 292579 | 1.85 | 0.64 | 19 | 1.50 | 0.63 | 13 | 1.23 | 0.65 | 11 |
| 278927 | 1.30 | 0.63 | 20 | 0.51 | 0.54 | 76 | 0.97 | 0.62 | 26 |
| $\sigma^2$ | 20.02 | | | 18.61 | | | 6.64 | | |

# Chapter 6

# Ignoring Overdispersion in Hierarchical Models: Possible Problems and Solutions

The combined model introduced in Chapter 5 can play a vital role in exploring the impact of ignoring important sources of variation, like bias. This was already seen when the model that ignores selection process was used to model data that clearly needed data augmentation. Using overdispersion as an important source of variation, this chapter investigates further, the impact of ignoring data enrichment when it is necessary. In addition to the impact on estimates and standard errors, the impact on type I error is also studied. We use the combined model as a tool for investigation, whose performance in situations where the distribution of either both, or one set of random effects is misspecified, is studied.

## 6.1 Introduction

The introduction of random effects to model correlated responses coming from the same subject, was a milestone contribution to the analysis of complex data Fisher (1918). Over the last decades, these hierarchical models have been applied in a multitude of areas like, item response theory (De Boeck and Wilson, 2004), toxicology (Molenberghs and Verbeke, 2005), survival analysis (Duchateau and Janssen, 2007) and non-linear mixed models (Davidian and Giltinan , 1995). Many of the models used in these fields fall under the umbrella of generalized linear mixed models (GLMM;

Agresti (2002)). Basically, GLMMs are used to account for the heterogeneity that arises from correlated measurements. However, in several applications, there may be additional sources of heterogeneity that can affect our inferences if ignored. Poisson longitudinal data are an archetypical example where heterogeneity may arise, not only from the repeated measurements, but also from additional overdispersionHinde and Demétrio (1998).

Several studies have explored the impact of misspecifying different aspects of GLMMs on the inferential procedures emanating from them. For instance, Agresti (2002) addressed the impact of omitting important confounding factors, Litière, Alonso and Molenberghs (2007) investigated the effect of misspecifying the random effects distribution and Ten Have and Tran (1999) assessed the impact of fitting an incomplete multilevel structure. All the previous research has clearly shown that misspecification may seriously affect our conclusions. Along these lines, in the present work we study the effect of ignoring overdispersion in hierarchical loglinear models.

Molenberghs *et al.* (2010) presented a model that deals with overdispersion by introducing an additional set of random effects in the classical Poisson loglinear mixed model. In the following, we shall refer to this model as the combined model. We will use this combined model as a framework to assess the impact of ignoring overdispersion via simulations. Essentially, we will study the impact of the misspecification on the consistency of the maximum likelihood estimators (MLE) and the Type I error rates.

Another important concern that arises when using this type of models is the distributional assumptions one needs to make for the random effects. Indeed Neuhaus, Hauck, and Kalbfleisch (1992) showed that misspecfying the random effect distribution in a logistic model may result in estimates that are asymptotically biased, though the bias is typically small. In a similar setting, Agresti, Caffo, and Ohman-Strickland (2004) found that misspecification of the random effects distribution may produce a loss of efficiency. Through extensive simulations Litière, Alonso and Molenberghs (2007) and Litière, Alonso, and Molenberghs (2008) investigated the impact of this misspecification on the consistency of the MLE, the power and Type I error rate of commonly used inferential procedures in GLMM. They observed that, although in most scenarios the estimates of many fixed effects were little affected, the estimates of variance components were severely biased and the power and Type I error rates were also gravely impacted.

Introducing an additional set of random effects obviously implies additional distributional assumptions for intrinsically unobservable latent variables. Therefore, in the present work, we will also explore the impact of misspecifying the random effects

distribution on both the classical Poisson loglinear mixed model and the combined model introduced by Molenberghs *et al.* (2010). Let $Y_{ij}$ represent the number of epileptic seizures patient $i$ experienced during week $j$ of the follow-up period. Further, let $t_{ij}$ be the time-point at which $Y_{ij}$ is measured, $t_{ij} = 1, 2, \ldots$ until at most 27. Following Molenberghs and Verbeke (2005) the next model was used to analyze the data,

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\kappa_{ij}), \\ \ln(\kappa_{ij}) &= \begin{cases} (\beta_0 + b_i) + \beta_1 t_{ij} & \text{if placebo,} \\ (\beta_2 + b_i) + \beta_3 t_{ij} & \text{if treated,} \end{cases} . \end{aligned} \tag{6.1}$$

where $b_i$ is assumed to follow $N(0, \sigma_b^2)$. The first part of Table 6.1 summarizes the main findings. The results indicate that the expected number of seizures significantly decreases over time in both the placebo and experimental group with p-values 0.0017 and 0.0067 respectively. Importantly, the rate of decrease was the same for both groups, i.e., no significant difference between the placebo and the new treatment was detected with a p-value of 0.7115. Obviously, the preceding results are conditional on the validity of the model used for the analysis. In that line, one relevant question is if the previous model suffices to account for all the variability present in the data and the impact of ignoring extra sources of variability on the inferences previously described. We will address this important issue at the end of the manuscript.

## 6.2 Combining Conjugate and Normal Random Effects

In this section, we will briefly introduce the model proposed by Molenberghs *et al.* (2010). To that effect, let us denote by $Y_{ij}$ the $j$th outcome in cluster $i = 1, \ldots, N$ with $j = 1, \ldots, n_i$. Furthermore, it will be assumed that, conditionally upon two $q$-$n_i$-dimensional vectors of random effects $\boldsymbol{b_i}$ and $\boldsymbol{\theta}_i$, the outcomes $Y_{ij}$ are independent with density function of the form

$$f_i(y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \theta_{ij}, \phi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\right\}, \tag{6.2}$$

where the conditional mean $\mu_{ij}^c$ is further modeled as

$$E(Y_{ij}|\boldsymbol{b_i}, \boldsymbol{\beta}, \theta_{ij}) = \mu_{ij}^c = \theta_{ij}\kappa_{ij}. \tag{6.3}$$

In the preceding expression the random variable $\theta_{ij} \sim \mathcal{G}_{ij}(\vartheta_{ij}, \sigma_{ij}^2)$ with $\vartheta_{ij}$ and $\sigma_{ij}^2$ denoting the mean and the variance of $\theta_{ij}$ respectively and $\kappa_{ij} = g(\boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b_i})$. Moreover, it will be typically assumed that $\boldsymbol{b_i} \sim N(\boldsymbol{0}, \boldsymbol{D})$.

It is convenient, but not strictly necessary, to consider that the two sets of random effects $\boldsymbol{\theta}_i$ and $\boldsymbol{b_i}$ are independent of each other. Regarding the components $\theta_{ij}$ of $\boldsymbol{\theta}_i$, three useful special cases result from assuming that: (1) they are independent; (2) they are correlated, implying that the collection of univariate distributions $\mathcal{G}_{ij}(\vartheta_{ij}, \sigma^2_{ij})$ needs to be replaced with a multivariate one; and (3) they are equal to each other, useful in applications with exchangeable outcomes $Y_{ij}$.

Obviously, parameterization (6.3) allows for random effects $\theta_{ij}$ capturing overdispersion, and formulated directly at the mean scale, whereas $\kappa_{ij}$ could be considered the GLMM component.

### 6.2.1 Combined Poisson Model for Count Data

From the general developments above, the Poisson model with gamma and normal random effects naturally follows. By way of overview, let us assemble all model elements

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\theta_{ij}\kappa_{ij}), \\ \kappa_{ij} &= \exp\left(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i}\right), \\ \boldsymbol{b_i} &\sim N(\boldsymbol{0}, \boldsymbol{D}), \\ \theta_{ij} &\sim \text{Gamma}(\lambda, \tau) \\ E(\boldsymbol{\theta}_i) &= E[(\theta_{i1}, \ldots, \theta_{in_i})'] = \boldsymbol{\vartheta}_i, \\ \text{var}(\boldsymbol{\theta}_i) &= \boldsymbol{\Sigma}_i. \end{aligned}$$

Essentially, this model has the same structure of the one by Booth *et al.* (2003) The $\theta_{ij}$ can be assumed independent and following a gamma distribution, producing, what we could term, a Poisson-gamma-normal model or, equivalently, a negative-binomial-normal model. This is natural in many cases in the sense that the $\boldsymbol{b_i}$ will induce association between repeated measurements, with then the $\theta_{ij}$ taking care of additional dispersion. In this case, $\boldsymbol{\Sigma}_i$ reduces to a diagonal matrix. Nevertheless, it is perfectly possible to allow for general covariance structures. When a fully distributional specification would be desired, then one could choose, for example, multivariate extensions of the gamma distribution .

The Poisson-gamma-normal model can be fitted following a two step procedure. In fact, integrating the previous conditional model over the gamma random effects, leaving the normal random effects untouched, leads to

$$f(y_{ij}|\boldsymbol{b_i}) = \begin{pmatrix} \lambda_j + y_{ij} - 1 \\ \lambda_j - 1 \end{pmatrix} \cdot \left(\frac{\tau_j}{1+\kappa_{ij}\tau_j}\right)^{y_{ij}} \cdot \left(\frac{1}{1+\kappa_{ij}\tau_j}\right)^{\lambda_j} \kappa_{ij}^{y_{ij}}, \qquad (6.4)$$

where $\kappa_{ij} = \exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{z}'_{ij}\boldsymbol{b_i})$. It is important to point out that in this approach the gamma random effects are assumed to be independent. Model (6.4) can now be easily fitted using maximization routines like the one implemented in PROC NLMIXED in SAS.

## 6.3 Simulation Studies

### 6.3.1 Impact of Ignoring Overdispersion

In this set of simulations, and using the Poisson-gamma-normal model as suitable framework to generate the data, we explore the impact of ignoring overdispersion on the parameter estimates and their standard errors. Mimicking the case study, longitudinal poisson responses $Y_{ij}$ were generated with mean $\theta_{ij}\kappa_{ij}$, where $\theta_{ij}$ was randomly sampled from $\Gamma(\lambda, \tau)$ and

$$\kappa_{ij} = \exp\left(\beta_0 + b_i + \beta_1 t_j + \beta_2 z_i + \beta_3 t_j z_i\right). \tag{6.5}$$

In the previous expression, $i = 1, ..., 500$ and $t_j = 1, 2, 3, 4, 5, 6$ denote the subject and the time of measurement, respectively. Moreover, $b_i \sim N(0, \sigma_b^2)$ and $z_i$ is a treatment group indicator variable taking values 0/1.

The data were generated using three sets of parameters, i) $\beta_0 = -2$, $\beta_1 = -0.5$, $\beta_2 = -3$ , $\beta_3 = 1$, $\lambda = 2$, ii) $\beta_0 = 0.1$, $\beta_1 = 0.2$, $\beta_2 = 0.3$ , $\beta_3 = 0.5$, $\lambda = 4$, and iii) like (ii) but with $\lambda = 0.5$. In order to vary the amount of overdispersion in the data, the $\theta_{ij}$s were sampled from three different gamma distributions: $\Gamma(4, 0.25)$, $\Gamma(2, 0.5)$ and $\Gamma(0.5, 1)$ and in all the cases $\sigma_b^2 = 4$. In total 500 data sets were generated in each setting and analyzed using the correct model

$$E(Y_{ij}|\theta_{ij}b_i) = \theta_{ij}\exp(\beta_0 + b_i + \beta_1 t_j + \beta_2 z_i + \beta_3 t_j z_i) \tag{6.6}$$

and a model that ignores the overdispersion, i.e., the misspecified model

$$E(Y_{ij}|b_i) = \exp(\beta_0 + b_i + \beta_1 t_j + \beta_2 z_i + \beta_3 t_j z_i). \tag{6.7}$$

The gamma distribution parameterization used in the above setting and all other settings to follow is,

$$f(y; \lambda, \tau) = \frac{y^{\lambda-1}\exp(-\frac{y}{\tau})}{\tau^{\lambda}\Gamma(\lambda)}$$

### 6.3.2 Impact on Incorrectly Assuming Overdispersion

In these simulations we studied the performance of model (6.6) when it is used to analyze data with no additional overdispersion. Basically, and using $\beta_0 = 0.1$, $\beta_1 =$

0.2, $\beta_2 = 0.3$ and $\beta_3 = 0.5$, $\sigma_b^2 = 4$ as values for the parameters, the data were generated using model (6.7) and latter analyzed using (6.6) and (6.7).

### 6.3.3 Impact of Misspecification of Random Effects

In order to study the impact of misspecifying the random effects distribution a new set of simulations was designed. Essentially, data were generated following the scheme presented in Section 6.3.1 using the following parameter values; $\beta_0 = 0.1$, $\beta_1 = 0.2$, $\beta_2 = 0.3$ and $\beta_3 = 0.5$, but three main variations were introduced: Firstly, the $b_i$s were generated using five different distributions: $N(0, 2)$, $\exp(\sqrt{0.5})$, $t_4$, $\Gamma(2, 0.5)$ and $\chi^2(1)$. Note that the parameters for the distributions were chosen such that $\text{var}(b_i) = 2$. The overdispersion random effects $\theta_{ij}$s were always sampled from a $\Gamma(4, 0.25)$. Secondly, the $\theta_{ij}$s were generated from $\Gamma(4, 0.25)$ and $\chi^2(2)$ and the $b_i$s were always sampled from $N(0, 2)$. Finally in the third scenario, the $b_i$s were sampled from the distributions aforementioned in the first setting and the $\theta_{ij}$s were generated from $\chi^2(2)$.

Eventually, Models 6.6 and 6.7 were fitted to the generated data with the distributional assumptions described in Section 6.2.1. The goal is to explore the impact of misspecifying every set of random effects separately or simultaneously on the inferences obtained from these models.

### 6.3.4 Type I Error

With these simulations we seek to establish whether the Type I error is preserved in the combined model and also when we have ignored the overdispersion. The data generation is as in section 6.3.1 but using $\beta_0 = -2$, $\beta_1 = -0.5$, $\beta_2 = 0$ and $\beta_3 = 1$, $\sigma_b^2$=4 $\lambda = 4$ as parameter values. The data were also fitted to both Models 6.6 and 6.7. For all situations, 500 datasets were generated and the sample size was 500.

## 6.4 Simulation Results

### 6.4.1 Ignoring Overdispersion

Table 6.2 summarizes the main findings of this study. Even though ignoring overdispersion may have a negative impact on the parameter estimates, particularly when the overdispersion distribution is highly skewed (when $\lambda = 0.5$), this impact was in general very mild. Indeed, for the covariates effect the relative bias never surpassed 7% and was frequently much smaller. Nevertheless, the intercept was severely affected

by the misspecification in some scenarios and the variance of the random effect was seriously biased in some settings as well.

Importantly, the standard errors for the Poisson-normal model were always underestimated, especially for the interaction parameter which, in many cases, provides the answer to the main research question.

Much as it is harmful to ignore overdispersion, it was found (results not shown) that there is no harm in fitting the combined model when there is no overdispersion. Actually, in this scenario, the parameter estimates and standard errors for the two models (combined and Poisson-normal) were very close. When fitting the combined model, the estimate of $\lambda$ was very large in all cases, which implies no overdispersion. This is a bit counterintuitive, but follows from the nature of the gamma distribution. Basically, the combined model converges to the Poisson-normal model when it is fitted to data that is not overdispersed.

### 6.4.2 Misspecification of Random Effects Distribution

The intangible nature of random effects makes the selection of their distributional assumptions quite arbitrary. In Table 6.3 we show the impact of misspecifying the distribution of $b_i$ on inferences based on both, the combine and the Poisson-normal model. In the combined model the covariate effects were rather robust with respect to the misspecification. Indeed, the relative bias was always smaller than 3% and the associated standard errors were close to those obtained when the random effect distribution was not misspecified. Nevertheless, like before, the intercept and the estimates of $\sigma_b^2$ were seriously affected in some scenarios.

We also studied the impact of misspecifying the distribution of $\theta_{ij}$ on inferences emanating from the combined model. The main results are presented in Table 6.4. Clearly, the impact of this misspecification is almost negligible and the estimates and the standard errors of all covariate effects are very close to the those obtained under the correctly specified model. The robustness of the combined models follows from additional flexibility due to the presence of the overdispersion. Of course, one is not certain that the posited form fully coincides with the true data generating mechanism. Therefore, one might consider further extensions, such as shape parameters that are not constant but a function of covariates. The key message, though, is that the step from a conventional GLMM to the combined model is a crucial one.

Turning to the Poisson-normal model, it is important to point out that in this setting this model misses two important features of the data: the presence of overdispersion and the real distribution of the random effect $b_i$. When $b_i$ was sampled from

a gamma and chisquare distribution, the Poisson-normal model never converged. For the t-distribution with 4 degrees of freedom ($t_4$) the parameter estimates for $\beta_2$ had a relative bias as large as 20%. Note that the $t$ distribution has wider tails than its normal counterpart. For large datasets, this can have non-negligible effect. This is considerable larger than the bias found when the model only missed the overdispersion in the data. Here again severe underestimation of the standard errors was observed.

Another important issue that emerged in this study was the low rate of convergence observed for both models. In fact, the rates of convergence for the rows of Table 6.3 from top to bottom were: 100%, 55.8%, 36.6%, 75.6% and 31.6% respectively.

Perhaps the most relevant situation for practical purposes is when both distributions are misspecified. Table 6.5 illustrates our findings in this scenario. In this situation, estimates associated to the variance components of both distributions are largely biased. Nonetheless, apart from the intercept, the other effects estimates are generally close to the true values. Therefore, even when the distributions associated with both sets of random effects are misspecified, the covariate effects can be reliably estimated.

### 6.4.3 Type I Error

In Section 6.4.1, we discussed the impact of ignoring overdispersion on parameter estimates and standard errors, in which we saw that there is a large impact on standard errors which can possibly lead to erroneous conclusions. In this section we studied the impact of ignoring overdispersion on the Type I error by simulating data with no treatment effect and fitting both the combined model and the Poisson-normal model with treatment effect. The pre-specified Type I error was 5%. For the combined model, out of the 500 datasets, the treatment effect was found to be significant in 27, which translates into 5.4% Type I error. On the other hand, out of the 331 datasets which converged for the Poisson-normal model, 61 found a significant treatment effect which represents 18.4% Type I error. In the best case scenario, if we assume that the models that failed to converge would not have detected a treatment effect, 61 out of 500 would translate into 12.2% which is still highly inflated. This finding is inline with the underestimation of standard errors by the Poisson-normal model as discussed in Section 6.4.1.

## 6.5 Re-Analyzing the Case Study

We also fitted the combined model to the data introduced in Section 2.2. The main findings are presented in the second part of Table 6.1. A number of remarks come into place here. Note first that, informally assessing the estimates of the parameters for the overdispersion random effects distribution, one can conclude that the overdispersion in these data should not be ignored. Complementing our findings is the observation that, the standard errors of the Poisson-normal model, especially the slope parameters are generally underestimated in the Poisson-normal. However, like before, the difference in the expected number of seizures between the experimental and placebo groups was not significant (p-value=0.2260) when the data were analyzed with the combined model. Note finally that, comparing the likelihood values, the combined model achieves a much better fitting of the observable data than the Poisson-normal model. In general, it is sensible to start with the combined model and, if the overdispersion effect is not significant, then one can switch to the more conventional GLMM.

## 6.6 Discussion

Blindly assuming that unobserved heterogeneity in repeated measurements data only comes from the correlation in the responses can be too restrictive and sometimes can lead to invalid conclusions. We have shown through simulations that ignoring overdispersion in count data can have dire consequences on estimation of some covariate effects and their standard errors, as well as, on the variance components and the Type I error rates. Importantly, we found that the Type I error rates were considerably inflated when overdispersion was ignored, implying that the probability of detecting a spurious effect increases. Remarkably, our findings are strikingly similar to those reported by (Litière, Alonso and Molenberghs, 2007; Litière, Alonso, and Molenberghs, 2008) when studying the impact of misspecifying the random effect distribution in a logistic model with a random intercept. It is interesting to see that two related but different types of misspecification, i.e., ignoring overdispersions and misspecifying the random effect distribution, may have very similar consequences.

Our simulations also indicate that the combined model may be a reasonable alternative in this situation. When the combined model is fitted to data that has no overdispersion, it converges to the Poisson-normal model and no numerical issues emerge in this situation. Furthermore, the model is rather robust to misspecification of the random effects distributions. All the previous characteristics seem to indicate

Table 6.1: *Epilepsy Study. Parameter estimates and standard errors for the regression coefficients in the Poisson-normal model, and the combined model. Estimation was done by maximum likelihood using numerical integration over the normal random effect, if present.*

| | | Poisson-normal | Combined |
|---|---|---|---|
| Effect | Parameter | Estimate (s.e.) | Estimate (s.e.) |
| Intercept placebo | $\beta_0$ | 0.8179 (0.1677) | 0.9112 (0.1755) |
| Slope placebo | $\beta_1$ | −0.0143 (0.0044) | −0.0248 (0.0077) |
| Intercept treatment | $\beta_2$ | 0.6475 (0.1701) | 0.6555 (0.1782) |
| Slope treatment | $\beta_3$ | −0.0120 (0.0043) | −0.0118 (0.0074) |
| overdispersion parameter | $\lambda$ | — | 2.4640 (0.2113) |
| overdispersion parameter | $\tau = 1/\lambda_1$ | — | 0.4059 (0.0348) |
| Variance of random intercepts | $\sigma_b^2$ | 1.1568 (0.1844) | 1.1289 (0.1850) |
| −2log-likelihood | | −6810 | −7664 |

that the combined model is a useful tool for the analysis of Poisson data with overdispersion and it there is no harm in always starting with the combined model in settings where overdispersion is suspected.

. .

.

.

Table 6.2: *Median of Parameter Estimates, {relative bias in %} and (standard errors) for simulations studying the impact of ignoring overdispersion. Data were generated from combined model with normal ($b_i$) and gamma ($\theta_{ij}$) random effects using different levels of skewness ($\lambda$). The data were analyzed with Comb=combined model, PN=Poisson-normal. TV denotes the true values.*

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma_b^2$ | $\lambda$ |
| TV | 0.1 | 0.2 | 0.3 | 0.5 | 4 | 4 |
| Comb | 0.102{2} | 0.200{0} | 0.288{4} | 0.500{0} | 3.990{0.25} | 4.020{0.5} |
| | (0.139) | (0.012) | (0.1926) | (0.0155) | (0.0694) | (0.176) |
| PN | 0.112{12} | 0.192{4} | 0.315{5.1} | 0.488{2.4} | 3.974{0.65} | |
| | (0.1315) | (0.0040) | (0.1828) | (0.0044) | (0.0684) | |
| TV | -2 | -0.5 | -3 | 1 | 4 | 2 |
| Comb | -2.004{0.2} | -0.499{0.2} | -2.982{0.6} | 1.003{0.3} | 4.149{3.7} | 2.002{0.1} |
| | (0.2450) | (0.0520) | (0.3404) | (0.0679) | (0.1365) | (0.4438) |
| PN | -2.074{3.7} | -0.496{0.8} | -2.964{1.2} | 0.995{0.6} | 4.140{3.5} | |
| | (0.2312) | (0.0395) | (0.2989) | (0.0498) | (0.1343) | |
| TV | 0.1 | 0.2 | 0.3 | 0.5 | 4 | 0.5 |
| Comb | 0.106{6} | 0.202{1} | 0.271{9.7} | 0.502{0.4} | 3.938{1.6} | 0.500{0} |
| | (0.1837) | (0.0300) | (0.2433) | (0.0403) | (0.0792) | (0.0190) |
| PN | -0.706{805} | 0.186{7} | 0.287{4.3} | 0.476{4.8} | 4.485{12.1} | |
| | (0.1441) | (0.0057) | (0.1976) | (0.0061) | (0.0763) | |

Table 6.3: *Median parameter estimate, {relative bias in %} and (standard errors) for simulations studying the impact of misspecfying the distribution of $b_i$. Data were generated using the combined model in which $b_i$ was sample from RE-dist. Further the Comb=combined and PN=Poisson-normal models were fitted assuming normality.*

| RE-dist. | | Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma_b^2$ | $\lambda$ |
| | | 0.1 | 0.2 | 0.3 | 0.5 | 2 | 4 |
| normal | comb | 0.105{5} | 0.199{0.5} | 0.291{3} | 0.501{0.2} | 1.989{0.55} | 4.027{0.68} |
| | | (0.1048) | (0.0123) | (0.1435) | (0.0156) | (0.0491) | (0.1803) |
| gamma | Comb | 2.078{1978} | 0.201{0.5} | 0.300{0} | 0.501{0.2} | 1.955{2.4} | 4.016{0.4} |
| | | (0.0956) | (0.0090) | (0.1343) | (0.0122) | (0.0453) | (0.1319) |
| chisquare | Comb | 1.038{938} | 0.200{0} | 0.307{2.3} | 0.500{0} | 1.797{10.5} | 3.992{0.2} |
| | | (0.0951) | (0.0105) | (0.1325) | (0.0136) | (0.0434) | (0.1534) |
| $t_4$ | Comb | 0.107{7} | 0.200{0} | 0.301{0.3} | 0.500{0} | 1.732{13.4} | 4.011{0.3} |
| | | (0.0996) | (0.0124) | (0.1360) | (0.0157) | (0.0460) | (0.1824) |
| $t_4$ | PN | 0.093{7} | 0.196{2} | 0.360{20} | 0.481{3.8} | 1.724{13.8} | |
| | | (0.0896) | (0.0060) | (0.1228) | (0.0064) | (0.0447) | |

Table 6.4: *Median parameter estimates, {relative bias in %} and (standard errors) for simulations studying the impact of misspecfying the distribution of $\theta_{ij}$. The $\theta_{ij}$s were sampled from Re-dist. The combined model was fitted assuming gamma distributed overdispersion random effects.*

| RE-dist. | Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma_b^2$ | $\lambda$ |
| | 0.1 | 0.2 | 0.3 | 0.5 | 2 | 4 |
| gamma | 0.105{5} | 0.199{0.5} | 0.291{3} | 0.501{0.2} | 1.989{0.55} | 4.027{0.68} |
| | (0.1048) | (0.0123) | (0.1435) | (0.0156) | (0.0491) | (0.1803) |
| chisquare | 0.784{684} | 0.201{0.5} | 0.290{0.3} | 0.501{0.2} | 1.992{2} | 1.002{75} |
| | (0.1483) | (0.0191) | (0.2069) | (0.0258) | (0.0700) | (0.0324) |

Table 6.5: *Median parameter estimates, {relative bias in %} and (standard errors) for simulations studying the impact of misspecfying both $\theta_{ij}$ and $b_i$ distributions. Data were generated from a combined model with $b_i$ sampled from Re-dist. and $\theta_{ij}$ sampled from a chisquare distribution. The data were analyzed with the combined model assuming normal and gamma random effects.*

| RE dist. | Parameter Estimates | | | | | |
|---|---|---|---|---|---|---|
| | $\beta_0$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\sigma_b^2$ | $\lambda$ |
| | 0.1 | 0.2 | 0.3 | 0.5 | 2 | 4 |
| normal | 0.105{5} | 0.199{0.5} | 0.291{3} | 0.501{0.2} | 1.989{0.55} | 4.027{0.68} |
| | (0.1048) | (0.0123) | (0.1435) | (0.0156) | (0.0491) | (0.1803) |
| $t_4$ | 0.804{704} | 0.200{0} | 0.276{8} | 0.501{0.2} | 1.714{14.3} | 1.001{75} |
| | (0.1112) | (0.0184) | (0.1546) | (0.0251) | (0.0485) | (0.0320) |
| chisquare | 1.751{1651} | 0.192{4} | 0.262{12.7} | 0.504{8} | 1.620{19} | 1.002{75} |
| | (0.009) | (0.0171) | (0.1438) | (0.0238) | (0.0418) | (0.0285) |

# Part II

# Flexible Methodology For Data With Random Sample Size

# Chapter 7

# Properties of Estimators in Exponential Family Settings With Observation-based Stopping Rules

Random sample size (RSS) trials are beneficial in medical research, although mean estimators used in fixed sample size (FSS) trials are not confidently used after such trials since they lose some good properties. In this chapter, we study the properties of mean estimators in an RSS setting from a new perspective, that leads to interesting findings.

## 7.1 Introduction

It is commonly known that statistical designs where the sample size is random pose challenges beyond the fixed sample-size case and that many findings are counter-intuitive. While this has been documented for situations where the sample size depends on the data, such as in sequential trials (Siegmund, 1978; Hughes and Pocock, 1988; Emerson and Fleming, 1990) or incomplete data (Little and Rubin, 2002), it is less widespread that such counterintuitive results apply even when the sample size is completely random (Grambsch, 1983; Barndorff-Nielsen and Cox, 1984), in the sense that both the collected and uncollected data have no relationship to the stochastic

mechanism governing the sample size. Liu and Hall (1999) provided a general theory for sequential studies, where the decision to either stop or continue the study at every interim look depends deterministically on the data collected up to that point. Molenberghs *et al.* (2013) generalized their results to the setting where the sample size may depend stochastically rather than deterministically on the observed data, a general setting that contains both sequential trials and completely random sample sizes (CRSS) as special cases. We refer to these three settings together as a stochastic stopping rule. Molenberghs *et al.* (2013) also discussed the related cases of incomplete longitudinal data, censored time-to-event data, joint modeling of survival and longitudinal data, and clustered data with random cluster sizes.

An important finding of Liu and Hall (1999) was that the commonly used sufficient statistics in deterministic stopping designs are incomplete, a property that will be defined in the next section. Molenberghs *et al.* (2013) generalized this to stochastic stopping rules and explore the implications of this for linear estimators based on the sample sum as well as on so-called marginal and conditional estimators. They found for stochastic stopping rules that the counterintuitive implications of a random sample size follows from two properties: (a) excluding the CRSS case, the sample size is *non-ancillary* given the sample sum; (b) the pair consisting of the accumulating sample sum and the sample size is an *incomplete* minimal sufficient statistic. These properties are defined in Section 7.2.

The work of Liu and Hall (1999) and Molenberghs *et al.* (2013) was confined to the special case of normally distributed outcomes. Further, Molenberghs *et al.* (2013) illustrated there developments with a random stopping rule of probit form. These specific choices allow for insightful expressions. The latter choice is not however necessary for deterministic stopping rules that can be cast in the form of continuation and stopping regions or, equivalently, the boundaries between them.

Extending the results in Liu and Hall (1999), Liu *et al.* (2006) presented a general deterministic stopping rule theory where the outcome follows a one-parameter exponential family, and also established incompleteness for this case. This implies, in particular, that there are infinitely many unbiased estimators, none with uniformly minimum variance. In this chapter, we show incompleteness in the one-parameter exponential family case, for a stochastic stopping rule, and derive explicit results for linear estimators as well as for marginal and conditional likelihood estimators. These general findings are then further illustrated in the normal case, making the connection to Molenberghs *et al.* (2013), and in the case where the outcomes are binary, and hence the sample sum is binomial.

Our findings are essentially as follows. The classical sample average is biased in

finite samples, though asymptotically unbiased for a broad classes of stopping rules. An unbiased estimator follows from the conditional likelihood, where the conditioning is on the (non-ancillary) sample size. Contrary to intuition, the conditional estimator has larger mean squared error than the ordinary sample average for sufficiently large sample size, the latter resulting from the joint likelihood, where 'joint' means a simultaneous model for the outcomes and the sample size. In some cases, the result holds for all sample sizes, large and small. Thus, the sample average is a valid and sensible estimator, contrary to some claims in the sequential-trial literature, for stochastic and deterministic stopping rules. The literature on sequential trials is indeed very large, with a relatively early review given by Whitehead (1999). Tsiatis, Rosner, and Mehta (1984) and Rosner and Tsiatis (1988) address precision estimation after group sequential trials. Emerson and Fleming (1990) propose estimators within an ordering paradigm. Much of this work is placed in a unifying framework by Liu and Hall (1999). A review can be found in Molenberghs *et al.* (2013).

The finite-sample bias in the sample average disappears only in the CRSS case. Even then, it is not unique in that a whole class of so-called generalized sample average estimators can be defined, all of which are unbiased. This enables us to show that the ordinary sample average is only asymptotically optimal. Indeed there is no uniformly optimal unbiased estimator in finite samples for most exponential-family members; the exponential distribution is a noteworthy exception.

The case of two possible sample sizes, $N = n$ and $N = 2n$ is simple yet generic, and will be adopted here. All developments can be generalized with ease to the setting with $L$ possible sample sizes and accrual numbers $n_1, \ldots, n_L$.

## 7.2 Notation, Basic Concepts, and Problem Formulation

As stated in the introduction, we consider a simple sequential trial, where $n$ measurements $Y_i$ are observed, after which a stochastic stopping rule is applied and, depending on the outcome, another set of $n$ measurements is or is not observed. Let $\boldsymbol{Y}$ be the $(2n \times 1)$ vector of outcomes that could be collected, with the sample sum denoted by $K$, and $N$ be the realized sample size, that is, $N = n$ or $N = 2n$. A joint model for the stochastic outcomes is

$$\begin{aligned} f(\boldsymbol{y}, N|\boldsymbol{\theta}, \boldsymbol{\psi}) &= f(\boldsymbol{y}|\boldsymbol{\theta}) \cdot f(N|\boldsymbol{y}, \boldsymbol{\psi}) && (7.1) \\ &= f(\boldsymbol{y}|N, \boldsymbol{\theta}, \boldsymbol{\psi}) \cdot f(N|\boldsymbol{\theta}, \boldsymbol{\psi}). && (7.2) \end{aligned}$$

The sample sum is denoted by $K$. If necessary, a subscript will indicate over which batch the sample is calculated. Molenberghs *et al.* (2013) noted the similarity with missing-data concepts, where (7.1) is a selection model factorization and (7.2) is a pattern-mixture factorization (Little and Rubin, 2002). In all cases, it is assumed that $f(N|\boldsymbol{y},\boldsymbol{\psi}) = f(N|\boldsymbol{y}^o,\boldsymbol{\psi})$ depends on observed outcomes only, and hence the sample size is determined by the first batch of observations $(Y_1,\ldots,Y_n)$. We may then write $f(N|K_n,\boldsymbol{\psi})$. This corresponds to the frequentist concept of missingness at random (Little and Rubin, 2002). In the limiting case of a deterministic stopping rule, $f(N|\boldsymbol{y},\boldsymbol{\psi})$ is degenerate and $f(N=n|\boldsymbol{y},\boldsymbol{\psi})$ equals 1 when $K_n \in \mathcal{S} \subset \mathbb{R}$ and 0 over its complement $\mathcal{C}$, with the reverse holding for $f(N=2n|\boldsymbol{y},\boldsymbol{\psi})$. The CRSS case follows by assuming $\boldsymbol{Y}$ and $N$ to be independent, meaning that both factorizations (7.1) and (7.2) trivially reduce to $f(\boldsymbol{y}|\boldsymbol{\theta})\cdot f(N|\boldsymbol{\psi})$.

In the stopping-rule case $\boldsymbol{\psi}$ is not estimable from the data and will be assumed to be specified by design. This is different for the other settings that can also be cast in terms of (7.1)–(7.2), such as incomplete longitudinal data, clusters of random size, censored time-to-event data, joint models for longitudinal and time-to-event data, and random measurement times settings, as noted by Molenberghs *et al.* (2013). In these cases, a subject-specific index $i$ needs to be introduced into (7.1)–(7.2) and $N$ needs to be replaced by the missing data indicators, censoring indicators, and so on.

### 7.2.1 Basic Concepts

In line with Molenberghs *et al.* (2013), we will review several fundamental concepts that are essential in what follows.

In agreement with Rubin (1976), we consider *ignorability*. For pure likelihood or Bayesian inferences, under missingness at random (MAR), inferences about $\boldsymbol{\theta}$ can be made using $f(\boldsymbol{y}_i^o|\boldsymbol{\theta})$ only, without the need for an explicit missing-data mechanism or, in our case, without the need for an explicit sample-size model. This is, provided the regularity condition of *separability* holds true, i.e., that the parameter space of $(\boldsymbol{\theta}',\boldsymbol{\psi}')'$ is the Cartesian product of their individual product spaces. In other words, this means that the sample size model does not contain information about the outcome model parameter. It implies that $N$ could then be considered ancillary in the sense of (Cox and Hinkley, 1974, pp. 32–35). We will see that this is true for CRSS, but not for the other situations. Excluding MNAR, ignorability can be violated in three ways. First, even in the likelihood and Bayesian frameworks and under MAR, ignorability does not apply in a non-separable situation. Second, frequentist inferences are not necessarily ignorable under MAR. Third, assuming MAR and separability hold and

we are in a likelihood or Bayesian framework, ignorability in the selection model decomposition (7.1) does not translate to the pattern-mixture model (7.2), as is clear from the presence of both $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ in both factors of (7.2). The latter statement is symmetric and could be made starting from a pattern-mixture view as well. The bottom line is that ignorability holds in at most one of these, except in the trivial MCAR setting, such as for CRSS.

There is a connection between ignorability and *ancillarity* (Cox and Hinkley, 1974). They define an ancillary statistic $T$ to be one that complements a minimally sufficient statistic $S$ such that, given $S$, $T$ does not contain information about the parameter of interest. Arguably the best known example is the sample size $T = n$ when estimating a mean, provided the sample size is fixed by design or the law governing it does not depend on the mean parameter to be estimated, as with CRSS. Counterexamples are the stochastic and deterministic stopping rules.

The crucial property for Liu and Hall (1999), Liu *et al.* (2006), Molenberghs *et al.* (2013), as well as for us here is that of *completeness* (Casella and Berger, 2001, pp. 285–286). A statistic $s(Y)$ of a random variable $Y$, with $Y$ belonging to a family $P_\mu$, is complete if, for every measurable function $g(\cdot)$, $E[g\{s(Y)\}] = 0$ for all $\mu$, implies that $P_\mu[g\{s(Y)\} = 0] = 1$ for all $\mu$. The relevance of completeness for us surfaces in two ways. First, from the Lehman-Scheffé theorem (Casella and Berger, 2001), if a statistic is unbiased, complete, and sufficient for some parameter $\mu$, then it is the best mean-unbiased estimator for $\mu$. The lack of this property in the stopping-rule case will manifest itself when studying generalized sample averages in Section 7.4. Second, completeness and ancillarity are connected through Basu's theorem (Basu, 1955; Casella and Berger, 2001, p. 287): a statistic both complete and sufficient is independent of any ancillary statistic.

### 7.2.2 General Model Formulation

Assume that we collect $n$ i.i.d. observations $Y_1, \ldots, Y_n$, with exponential family density

$$f_\mu(y) = h(y) \exp\{\mu y - a(\mu)\}, \tag{7.3}$$

where $\mu$ is the natural parameter, $a(\mu)$ the mean generating function, and $h(y)$ a normalizing constant. Assume a stochastic stopping rule

$$\pi(N = n|k_n) = F(k_n|\boldsymbol{\psi}) = F(k_n), \tag{7.4}$$

with $K_n = \sum_{i=1}^n Y_i$. The form for (7.4) is left unspecified at this time. The CRSS setting follows as $F(k_n) \equiv F$, a constant. Likewise, when $F(\cdot)$ is degenerate, a deterministic stopping rule ensues. When the trial is not stopped, a further $n$ observations

$Y_{n+1}, \ldots, Y_{2n}$ are collected, also with density (7.3). The inferential goal is to estimate $\mu$ or a function of this, such as the population mean $\mu$. From the exponential-family structure, the density of $K_n$ can be expressed

$$f_{n,\mu}(k) = h_n(k) \exp\{\mu k - na(\mu)\}. \tag{7.5}$$

When no ambiguity can arise, the subscript $n$ may be dropped from $K_n$. Because the density integrates to 1, it trivially follows that

$$e^{na(\mu)} = \int h_n(k) e^{\mu k} dk = \mathcal{L}\{h_n(k)\}. \tag{7.6}$$

While expression (7.6) is well known to be a Laplace transformation, it is useful to state it explicitly in preparation of the derivations in Section 7.3.

When, in addition, the conditional probability of stopping an exponential family form is chosen, e.g.,

$$F(k_n) = F(k) = \int_{z=-\infty}^{z=A(k)} \widetilde{f}_1(z) dz, \tag{7.7}$$

then an appealing form for the marginal stopping probability can be derived. Here $\widetilde{f}_1(z)$ can be seen as an exponential family member, underlying the stopping process. When the outcomes $Y$ and hence $K$ do not range over the entire real line, the lower integration limit in (7.7) should be adjusted accordingly, and the function $A(k)$ should be chosen so as to obey the range restrictions. It is convenient to assume that $\widetilde{f}_1(z)$ has no free parameters; should there be the need for such, then they can be absorbed into $A(k)$. Hence, we can write

$$\widetilde{f}_1(z) = \widetilde{h}_1(z) \exp\{-\widetilde{a}(0)\}. \tag{7.8}$$

Using (7.5) and (7.8), the marginal stopping probability becomes:

$$\begin{aligned} P(N=n) &= \int_{k=-\infty}^{k=+\infty} \int_{z=-\infty}^{z=A(k)} f_{n,\mu}(k) \widetilde{f}_1(z) dz\, dk \\ &= \exp\{-na(\mu) - \widetilde{a}(0)\} \int_{k=-\infty}^{k=+\infty} h_n(k) \left[ \int_{z=-\infty}^{z=A(k)} \widetilde{h}_1(z) dz \right] e^{\mu k} dk \\ &= \exp\{-na(\mu) - \widetilde{a}(0)\} \mathcal{L}\{H_1(A(k)) \cdot h_n(k)\}, \end{aligned} \tag{7.9}$$

where

$$H_1(t) = \int_{z=-\infty}^{z=t} \widetilde{h}_1(z) dz.$$

In the special case of a CRSS, $A(k) \equiv A$ and (7.9) reduces to

$$\begin{aligned} f(N=n) &= \exp\{-na(\mu) - \widetilde{a}(0)\} H_1(A) \mathcal{L}\{h_n(k)\} \\ &= \exp\{-na(\mu) - \widetilde{a}(0)\} H_1(A) e^{na(\mu)} = \int_{k=-\infty}^{A} \widetilde{f}_1(k) dk. \end{aligned}$$

In our two special cases, (7.3) will be chosen as standard normal and Bernoulli, respectively. In the first of these, in concordance with Molenberghs *et al.* (2013), (7.4) will be assumed to be of probit form:

$$F(k) = \Phi\left(\alpha + \beta\frac{k}{n}\right). \tag{7.10}$$

In the binary case, we will generally leave (7.4) unspecified, but for some developments it is useful to consider an explicit example, for which we will resort to the beta distribution, i.e.,

$$\widetilde{f}_1(z) = \frac{z^{\alpha-1}(1-z)^{\beta-1}}{B(\alpha,\beta)}, \tag{7.11}$$

with $B(\cdot,\cdot)$ the beta function. It is convenient to choose integer values, for illustrative purposes: $\alpha = p+1$, $\beta = q+1$, with $p$ and $q$ integers, changing (7.11) to:

$$\widetilde{f}_1(z) = (p+q+1)\left(\begin{array}{c} p+q \\ p \end{array}\right) z^p(1-z)^q. \tag{7.12}$$

Choosing (7.12) leads to the conditional stopping probability:

$$F(k) = (p+q+1)\left(\begin{array}{c} p+q \\ p \end{array}\right)\sum_{\ell=0}^{q}\frac{(-1)^\ell}{p+\ell+1}\left(\begin{array}{c} q \\ \ell \end{array}\right)A(k)^{p+\ell+1}. \tag{7.13}$$

It is instructive to consider some special cases of this. When $p = q = 0$, (7.12) reduces to the uniform distribution on the unit interval, and it immediately follows that $F(k) = A(k)$. When $p = 1$ and $q = 0$, we find $F(k) = A(k)^2$. As a third and last instance, when $p = q = 1$, $F(k) = 3A(k)^2 - 2A(k)^3$.

A useful function is $A(k) = k/n$, implying that stopping is certain when $K = n$ and continuation is certain when $K = 0$, while for $0 < K < n$ stopping is probabilistic. The actual probability in these cases depends on the choice for $p$ and $q$.

These choices are made to illustrate our general developments and our emphasis is not on, say, designing a particular trial. However, the class of beta-based stopping rules, for example, potentially leads to rich families of stopping rules and spending functions (Whitehead, 1997; Jennison and Turnbull, 2000).

## 7.3 Incomplete Sufficient Statistics

### 7.3.1 The General Case

We now consider the role of completeness in this setting, building upon the work of Liu and Hall (1999), Liu *et al.* (2006), and Molenberghs *et al.* (2013). A sufficient

statistic for this setting is $(K, N)$. In line with the developments in the above papers, the joint distribution for $(K, N)$ is:

$$p(k,n) = f_n(k) \cdot F(k), \tag{7.14}$$

$$p(k,2n) = f_{2n}(k) - \int f_n(z) f_n(k-z) F(z) dz. \tag{7.15}$$

When the stopping rule leads to range restrictions in the sense of Lehman (1950), it is known that the sufficient statistic is complete. Hence, for the rest of this section, we assume their necessary and sufficient conditions do not hold. It is known that these conditions do not hold for the normal distribution, in contrast to classes of stopping rules for the Poisson and binomial distributions, for example.

Assume now that a function $g(K, N)$ exists such that its expectation is zero for all values of the parameter and further that integrands are not zero almost everywhere over their integration ranges. Such a function must satisfy:

$$\int g(k,n) f_n(k) F(k) dk + \int g(k,2n) f_{2n}(k) dk - \iint g(k,2n) f_n(z) f_n(k-z) F(z) dk\, dz = 0. \tag{7.16}$$

Substituting the general exponential form (7.5) into (7.17), and using (7.6), leads to

$$\int g(k,2n) h_{2n}(k) e^{\mu k} dk - \int g(k,2n) \left[ \int h_n(z) h_n(k-z) F(z) dz \right] e^{\mu k} dk = \int h_n(k) e^{\mu k} \cdot \int g(k,n) h_n(k) F(k) e^{\mu k} dk \tag{7.17}$$

Because the left hand side of (7.18) is a convolution, and using the uniqueness of the Laplace transform, we find:

$$g(k,2n) = -\frac{\int g(z,n) h_n(z) h_n(k-z) F(z) dz}{h_{2n}(k) - \int h_n(z) h_n(k-z) F(z) dz}. \tag{7.18}$$

Hence, when $g(k, n)$ is chosen arbitrarily, (7.18) prescribes the choice for $g(k, 2n)$ which leads to a counterexample to completeness, hence establishing incompleteness.

For the CRSS case, when $F(k) \equiv F$, a constant, and also choosing $g(k, n) = c$, a constant, it follows that

$$g(k,2n) = \frac{-F}{1-F} \cdot c.$$

In the limiting case of a deterministic stopping rule, $F(z) = 1$ over the stopping region $\mathcal{S}$ and 0 over its set complement $\mathcal{C}$. It then follows that (7.14)–(7.15) reduce

to:

$$p(k,n) = f_n(k) \cdot I(k \in \mathcal{S}), \tag{7.19}$$

$$p(k,2n) = \int_{\mathcal{C}} f_n(z) f_n(k-z) dz. \tag{7.20}$$

For the deterministic case, (7.18) becomes:

$$g(k,2n) = -\frac{\int_{\mathcal{S}} g(z,n) h_n(z) h_n(k-z) dz}{h_{2n}(k) - \int_{\mathcal{S}} h_n(z) h_n(k-z) dz} = -\frac{\int_{\mathcal{S}} g(z,n) h_n(z) h_n(k-z) dz}{\int_{\mathcal{C}} h_n(z) h_n(k-z) dz}. \tag{7.21}$$

Expression (7.21) follows from the fact that, in the deterministic case, $F(k) = 1$ over the stopping region $\mathcal{S}$ and 0 elsewhere. The transition from one denominator to the other follows from observing that the convolution of $f_n(k)$ with itself produces $f_{2n}(k)$, and then replacing all of these by their explicit exponential-family form (7.5). Alternatively, it is easy to show that (7.21) follows immediately from the definition of a function $G(K,N)$ and (7.19)–(7.20).

The implication of these findings is that whenever they hold, the Lehmann-Scheffé theorem cannot be applied (see Section 7.2). It follows that a best mean-unbiased estimator does not necessarily exist for the average. In the next section, it will be shown that this is indeed the case for many, but not all outcome distributions and stopping rules, given that, for example, the exponential distribution does admit a uniform optimum. It will be shown that no optimum exists for the normal case, in line with Molenberghs *et al.* (2013), and neither for the Bernoulli and Poisson cases, for a wide class of stopping rules.

### 7.3.2 The Normal Case

Following Molenberghs *et al.* (2013), consider the outcome to be standard normal with mean $\mu$ and let stopping be governed by (7.10). They derived from first principles that the marginal probability of stopping is:

$$P(N=n) = \Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1+\beta^2/n}}\right). \tag{7.22}$$

This expression also follows as a special case of (7.9) by choosing (7.10) as the stopping rule, i.e., $\widetilde{f}_1(z)$ as the standard normal density and $A(k) = \alpha + \beta k/n$, and further $f_{n,\mu=\mu} = \varphi_{\mu,n}(k)$, where $\varphi_{\mu,s}(k)$ is the normal density with mean $\mu$ and variance $s$. Details of this derivation are provided in Appendix B.

Clearly, (7.22) depends on $\mu$, implying that this pattern-mixture formulation is non-separable. In contrast, although the observed data are present in the conditional stopping probability, $\mu$ is not, implying separability in the selection model formulation.

In this case (7.14)–(7.15) takes the form

$$p_\mu(N,k) = p_0(N,k) \cdot \exp\left(k\mu - \frac{1}{2}n\mu^2\right) \tag{7.23}$$

with

$$p_0(n,k) = \phi_n(k) \cdot \Phi\left(\alpha + \frac{\beta}{n}k\right), \tag{7.24}$$

$$p_0(2n,k) = \phi_{2n}(k) \cdot \left[1 - \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)\right]. \tag{7.25}$$

Here, $\phi_s(k)$ is the normal density with mean 0 and variance $s$. Expression (7.25) is more explicit than (7.15), making use of the fact that the outcome densities are normal and the stopping probability is written as a normal cumulative distribution function. The derivation can be found in Molenberghs *et al.* (2013). Based on the fact that integrating the joint densities specified by (7.23)–(7.25) over $K$ and summing over $N$ should be equal to one, leads to the identity:

$$\int \varphi_{\mu,n}(k) \cdot \Phi\left(\alpha + \frac{\beta}{n}k\right) dk = \int \varphi_{\mu,2n}(k) \cdot \Phi\left(\frac{\alpha + \frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right) dk. \tag{7.26}$$

In Section 7.4.1, (7.26) will be derived in general.

The specific form of condition (7.18) is:

$$g(k,2n) \cdot p_0(2n,k) = -\int \phi_n(k-z) \cdot g(z,n) \cdot \phi_n(z) \cdot \Phi\left(\alpha + \frac{\beta}{n}z\right) dz. \tag{7.27}$$

In the CRSS case, (7.24)–(7.25) reduce to:

$$p_0(n,k) = \phi_n(k) \cdot \Phi, \tag{7.28}$$

$$p_0(2n,k) = \phi_{2n}(k) \cdot (1-\Phi), \tag{7.29}$$

where $\Phi \equiv \Phi(\alpha)$. Then here, as in the general case, (7.27) simplifies and leads to an explicit solution for a number of cases, especially when $g(k,n)$ is chosen to be a constant.

In addition, for this case, other explicit examples can be constructed, even when $\beta \neq 0$. We reproduce the two examples of Molenberghs *et al.* (2013).

**Example 1.** For the first of two examples, choose

$$g(k,n) = \widetilde{\lambda}, \tag{7.30}$$

an arbitrary constant. Then it immediately follows from (7.27) that

$$g(k,2n) = -\widetilde{\lambda} \cdot \frac{\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}{1-\Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}. \tag{7.31}$$

When $\beta = 0$, then the right hand side of (7.31) is constant and we can set $\widetilde{\lambda} = \lambda(1-\Phi)$, leading to $g(k,n) = \lambda(1-\Phi)$ and $g(k,2n) = -\lambda\Phi$.

Such $g(k,N)$ functions lead to entire classes of estimators. To see this, assume that an estimator for $\mu$ is available, $\widehat{\mu}$, say. For example, $\widehat{\mu}$ could be the sample average

$$\widehat{\mu} = \frac{1}{N}K. \tag{7.32}$$

We can then construct a class of estimators derived there from.

Applying this to our example and choosing (7.30) and (7.31) for the special case of $\beta = 0$ leads to the following class of estimators:

$$\widehat{\mu}_\lambda = \overline{\mu} + \lambda \cdot \left[(1-\Phi)I(N=n) - \Phi I(N=2n)\right]. \tag{7.33}$$

It follows directly from the construction of $g(k,N)$ that $E(\overline{\mu}) = E(\widehat{\mu}_\lambda)$ and hence, if $\overline{\mu}$ is unbiased, then so is $\widehat{\mu}_\lambda$.

For the variance of (7.33), we obtain $\mathrm{var}(\widehat{\mu}_\lambda) = \mathrm{var}(\overline{\mu}) + \lambda^2\Phi(1-\Phi)$ which, within this class, is minimal for $\lambda = 0$. Hence, for $\beta = 0$, i.e., the CRSS case, the original estimator is more efficient than any member of the new class. This will change when $\beta \neq 0$. We also need to consider the basic estimator itself, e.g., either (7.32) or (7.51). before moving on to this, we first complete the second example.

When $\beta \to +\infty$, (7.24)–(7.25) reduces to

$$p_0(n,k) = \phi_n(k) \cdot I(k>0), \tag{7.34}$$

$$p_0(2n,k) = \int_{k=-\infty}^{k=0} f_0(z,n) \cdot \phi_n(k-z)\, dz. \tag{7.35}$$

In this case, the $G(K,N)$ functions will take a particular form.

With $g(k,n)$ as in (7.30), when $\beta \to +\infty$ (7.31) becomes

$$g(k,2n) = -\widetilde{\lambda} \cdot \frac{\Phi\left[(2n)^{-1/2}k\right]}{1-\Phi\left[(2n)^{-1/2}k\right]}. \tag{7.36}$$

To see that the considerations particular to the above example are not unique, we consider a second one.

**Example 2.** Choose

$$g(k,n) = \frac{\lambda}{\Phi\left(\alpha + \frac{\beta}{n}k\right)}, \tag{7.37}$$

with $\lambda$ a given constant, then

$$g(k,2n) = -\frac{\lambda}{1 - \Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)}. \tag{7.38}$$

Choosing (7.37) and (7.38) now produces the estimator

$$\widetilde{\mu}_\lambda = \overline{\mu} + \lambda \cdot \left[ \frac{I(N=n)}{\Phi\left(\alpha + \frac{\beta}{n}k\right)} - \frac{I(n=2n)}{1 - \Phi\left(\frac{\alpha+\frac{\beta k}{2n}}{\sqrt{\frac{2n+\beta^2}{2n}}},\right)} \right]. \tag{7.39}$$

When now taking the limit $\beta \to +\infty$, (7.37)–(7.38) become:

$$g(k,n) = \lambda, \quad \text{defined for } k > 0, \tag{7.40}$$

$$g(k,2n) = -\frac{\lambda}{1 - \Phi\left[(2n)^{-1/2}k\right]}. \tag{7.41}$$

The fact that the function $g(k,n)$ in (7.40) is undefined over the negative real numbers is unproblematic, because the stopping region is confined to the non-negative half line.

### 7.3.3 The Binary Case

While the binary case follows from the general considerations given in Section 7.3.1, it is insightful to examine this outcome type in some detail; here, integration is replaced by summation. Let the Bernoulli probability be $\pi$. The sample sum $K$ then follows a $\text{Bin}(\pi, N)$ distribution and

$$f_{N,\pi}(k) = \left( \begin{array}{c} N \\ k \end{array} \right) \pi^k (1-\pi)^{N-k}. \tag{7.42}$$

For now, as in the general case, we leave $F(k)$ unspecified. The joint distribution of $(K, N)$ now takes the form

$$p(k,n) = \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} F(k), \tag{7.43}$$

$$p(k,2n) = \pi^k (1-\pi)^{2n-k} \left[ \left( \begin{array}{c} 2n \\ k \end{array} \right) - H(k) \right] = \pi^k (1-\pi)^{2n-k} \widetilde{H}(k), \tag{7.44}$$

where

$$H(k) = \sum_{z=0\vee(k-n)}^{k\wedge n} \left( \begin{array}{c} n \\ z \end{array} \right) \left( \begin{array}{c} n \\ k-z \end{array} \right) F(z), \tag{7.45}$$

the meaning of $\widetilde{H}(k)$ is obvious, $a \vee b = \max(a, b)$, and $a \wedge b = \min(a, b)$.

When stopping rule (7.13) is chosen, (7.43) becomes:

$$\begin{aligned} p(k,n) = & \\ & \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} (p+q+1) \\ & \left( \begin{array}{c} p+q \\ p \end{array} \right) \sum_{\ell=0}^{q} \frac{(-1)^\ell}{p+\ell+1} \left( \begin{array}{c} q \\ \ell \end{array} \right) A(k)^{p+\ell+1}. \end{aligned} \tag{7.46}$$

The marginal stopping probability can be derived by summing (7.47) over $k$ but is generally unwieldy. In the particular case that $p = q = 0$ and $A(k) = k/n$, we find

$$p(k,n) = \frac{k}{n} \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k}, \tag{7.47}$$

$$p(k,2n) = \frac{2n-k}{2n} \left( \begin{array}{c} 2n \\ k \end{array} \right) \pi^k (1-\pi)^{2n-k}. \tag{7.48}$$

While the derivation of (7.47) is obvious, that of (7.48) is less straightforward and details are given in Appendix B. From (7.47), we deduce immediately that

$$P(N=n) = \frac{1}{n} \sum_{k=0}^{n} k \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} = \pi.$$

In other words, this particular choice of conditional stopping rule produces essentially the simplest possible marginal stopping probability that depends on the parameter $\pi$ that governs the outcomes.

The condition for the existence of a non-trivial function $g(K, N)$ with expectation zero for all $\pi$ is a discrete version of (7.17) and reads:

$$\sum_{k=0}^{n} g(k,n) F(k) \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} + \sum_{k=0}^{2n} g(k,2n) \widetilde{H}(k) \pi^k (1-\pi)^{2n-k} = 0. \tag{7.49}$$

Writing $\gamma = \pi/(1-\pi)$, (7.49) becomes

$$\sum_{k=0}^{n} g(k,n) F(k) \left( \begin{array}{c} n \\ k \end{array} \right) \gamma^k + \sum_{k=0}^{2n} g(k,2n) \widetilde{H}(k) (1-\pi)^n \gamma^k = 0.. \tag{7.50}$$

Using the discrete-data version of (7.6), i.e.,

$$(1-\pi)^{-n} = \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) \gamma^k,$$

it follows that

$$-\sum_{k=0}^{2n} \left[ \sum_{z=0\vee(k-n)}^{k\wedge n} \left( \begin{array}{c} n \\ z \end{array} \right) \left( \begin{array}{c} n \\ k-z \end{array} \right) g(z,n)F(z) \right] \mu^k = \sum_{k=0}^{2n} g(k,2n)\widetilde{H}(k)\mu^k.$$

Owing to equality of polynomial coefficients, we find:

$$g(k,2n) = -\frac{\displaystyle\sum_{z=0\vee(k-n)}^{k\wedge n} \left( \begin{array}{c} n \\ z \end{array} \right) \left( \begin{array}{c} n \\ k-z \end{array} \right) g(z,n)F(z)}{\widetilde{H}(k)},$$

the discrete-data version of (7.18).

## 7.4 Generalized Sample Averages

### 7.4.1 The General Case

To underscore the impact of incompleteness of the statistics $(K,N)$, Molenberghs *et al.* (2013) generalized the sample average (7.32) to

$$\overline{\mu} = \frac{K}{N}\cdot[c\cdot I(N=n) + d\cdot I(N=2n)] = K\cdot\left[\frac{c\cdot I(N=n)}{n} + \frac{d\cdot I(N=2n)}{2n}\right], \quad (7.51)$$

for some constants $c$ and $d$. We will refer to it as the *generalized sample average* (GSA). The ordinary sample average follows as $c=d=1$. In this section, (7.51) will be considered from a general exponential-family perspective. Sections 7.4.2 and 7.4.3 bring out some further specifics for the normal and Bernoulli cases, respectively.

From (7.5), the mean follows as $\mu = \partial a(\mu)/\partial\mu$. The expectation is:

$$\begin{aligned} E(\overline{\mu}) = \frac{c}{n}\int k f_n(k)F(k)dk + \frac{d}{2n}\int k f_{2n}(k)dk \\ -\frac{d}{2n}\iint k f_n(z)f_n(k-z)F(z)dk\,dz. \end{aligned} \quad (7.52)$$

This form can be simplified. We will derive two identities that are useful here and in what follows. Because integrating (7.14)–(7.15) over $K$ and summing over $N$ should lead to unity, it follows that

$$\int f_n(k)F(k)dk = \iint f_n(z)f_n(k-z)F(z)dk\,dz. \quad (7.53)$$

This equation obviously also follows from first principles. Likewise, we have that

$$\begin{aligned}\iint k f_n(z) f_n(k-z) F(z) dk\, dz &= \int_z f_n(z) F(z) \left[ \int_k k f_n(k-z) dk \right] dz \\ &= \int_z f_n(z) F(z) \left[ n\mu + z \right] dz \\ &= n\mu A_n(\mu) + B_n(\mu), \end{aligned} \tag{7.54}$$

where

$$A_n(\mu) = \int f_n(k) F(k) dk, \qquad B_n(\mu) = \int k f_n(k) F(k) dk. \tag{7.55}$$

Using (7.54), we can rewrite (7.53) as

$$E(\overline{\mu}) = d \cdot \mu + \frac{2c-d}{2n} B_n(\mu) - \frac{d \cdot \mu}{2} A_n(\mu) \tag{7.56}$$

$$= d \cdot \mu + \frac{1}{2} \left[ (2c-d) E\left( \frac{K}{N} \middle| N = n \right) - d \cdot \mu \right] \cdot P(N = n). \tag{7.57}$$

While obvious, it is useful to spell out (7.56)–(7.57) for the ordinary sample average:

$$E(\overline{\mu}) = \mu + \frac{1}{2n} B_n(\mu) - \frac{\mu}{2} A_n(\mu) \tag{7.58}$$

$$= \mu + \frac{1}{2} \left[ E\left( \frac{K}{N} \middle| N = n \right) - \mu \right] \cdot P(N = n). \tag{7.59}$$

It is very intuitive that the bias in the sample average is a simple function of the difference between conditional and marginal expectation of $K/N$ on the one hand, and the probability of stopping on the other.

The specific form of (7.53) will depend on both the exponential family member considered and the form of the stopping rule. In general, the expectation may be a non-linear function of $\mu$ and hence there may be no constants $c$ and $d$ for which the expectation is $\mu$. Hence, in many situations, all linear estimators of the form (7.51) may be biased. Examples are given in Sections 7.4.2 and 7.4.3.

We now turn to the asymptotic behavior of the GSA, i.e., the case where $n \to +\infty$. Because $K$ converges to a $N(n\mu, n\sigma^2)$ variable, and using a first-order Taylor series expansion $F(k) \approx F(n\mu) + F'(n\mu)(k - n\mu)$, we find from first principles:

$$A_n(\mu) \approx F(n\mu), \tag{7.60}$$

$$B_n(\mu) \approx n\mu F(n\mu) + n\sigma^2 F'(n\mu). \tag{7.61}$$

Using (7.60) and (7.61), (7.56) converges to:

$$E(\overline{\mu}) \overset{n \to +\infty}{\longrightarrow} d \cdot \mu + (c-d) \cdot \mu \lim_{n \to +\infty} F(n\mu) + \frac{2c-d}{2} \sigma^2 \lim_{n \to +\infty} F'(n\mu). \tag{7.62}$$

In particular, for the ordinary sample average:

$$E(\overline{\mu}) \overset{n\to+\infty}{\longrightarrow} \mu + \frac{1}{2}\sigma^2 \lim_{n\to+\infty} F'(n\mu). \tag{7.63}$$

In Section 7.4.2, we we will see that (7.62) is finite and, moreover, (7.63) equals $\mu$. Sufficient conditions for this to hold in general can be given. Assume that $F(\cdot)$ is a continuously differentiable function that depends on $k$ as a function of $k/n$. To emphasize this, write

$$F(k) = F(\eta(k/n)). \tag{7.64}$$

Then $F(n\mu) = F(\eta(\mu)$, independent of $n$ and $F'(n\mu) = \eta'(\mu)F'(\eta(\mu))/n$, which depends on $n$ only through the factor $n^{-1}$ and hence converges to zero. More generally, a stopping rule that satisfies $F'(n\mu) \overset{n\to+\infty}{\longrightarrow} 0$ ensures that the sample average is asymptotically unbiased.

For a GSA to be asymptotically unbiased, (7.62) should equal $\mu$. Assume that the third term on the right hand side of (7.62) is zero and $F(n\mu) = \widetilde{F}(\mu)$ does not depend on $n$. The GSA is unbiased if $d + (c-d)\widetilde{F}(\mu) = 1$ for all values of $\mu$ (note that, when $\mu = 0$, the limit is trivially equal to zero). This equation can be satisfied if $\widetilde{F}(\mu)$ is constant, i.e., in the CRSS case to be discussed next. Otherwise, the equation can be satisfied only for $c = d = 1$, i.e., the ordinary sample average.

For the GSA to be unbiased in the finite-sample case, (7.56) needs to equal $\mu$, leading to the requirement:

$$d = \frac{2\mu - 2c\widetilde{\mu}P(N=n)}{2\mu - (\widetilde{\mu}+\mu)P(N=n)}, \tag{7.65}$$

with $\widetilde{\mu} = E(K/N|N=n)$. Evidently, this is a function of $\mu$ in the non-CRSS case and hence no uniformly unbiased estimator exists. Further, unless in the CRSS case, the ordinary sample average never satisfies (7.65) because this would imply that $\widetilde{\mu} = \mu$ and hence the stopping probability would be independent of $\mu$.

In the specific case of a CRSS, the constant $F$ is taken out of the integrals on the right hand side of (7.56) and we easily find:

$$E_{\text{CRSS}}(\overline{\mu}) = [cF + d(1-F)]\mu, \tag{7.66}$$

which is unbiased if and only if

$$d = \frac{1-cF}{1-F}. \tag{7.67}$$

An obvious solution is $c = d = 1$, the sample average, next to an infinite number of unbiased linear estimators of the type (7.51). Note that (7.67) follows from (7.65) upon observing that in the CRSS case $\mu = \widetilde{\mu}$ and $P(N=n) = F$.

In addition to studying the overall expectation of the GSA, it is of interest to consider the conditional expectations. These are:

$$E(\overline{\mu}|N=n) = \frac{c}{n}\cdot\frac{B_n(\mu)}{A_n(\mu)}, \tag{7.68}$$

$$E(\overline{\mu}|N=2n) = \frac{d}{2n}\cdot\frac{2n\mu - n\mu A_n(\mu) - B_n(\mu)}{1-A_n(\mu)}. \tag{7.69}$$

The ordinary sample average versions follow by setting $c=d=1$ in (7.68)–(7.69).

The asymptotic behavior of (7.68)–(7.69), follows from applying (7.60) and (7.61):

$$E(\overline{\mu}|N=n) \overset{n\to+\infty}{\longrightarrow} c\left(\mu+\sigma^2\lim_{n\to+\infty}\frac{F'(n\mu)}{F(n\mu)}\right), \tag{7.70}$$

$$E(\overline{\mu}|N=2n) \overset{n\to+\infty}{\longrightarrow} d\left(\mu-\frac{1}{2}\sigma^2\lim_{n\to+\infty}\frac{F'(n\mu)}{1-F(n\mu)}.\right). \tag{7.71}$$

For the ordinary sample average, when $F'(n\mu)$ converges to zero, the conditional expectations converge to $\mu$. In case the limits in (7.70) and (7.71) differ from zero, there is a choice for $c$ and $d$ that produces conditional expectations equal to $\mu$: $c=\mu/[\mu+\sigma^2Q_1(\mu)]$ and $d=\mu/[\mu-0.5\sigma^2Q_2(\mu)]$, with obvious notation. Evidently, these are not uniform and therefore not useful in practice. These values for $c$ and $d$ lie at different sides of unity. We will return to the implications of limiting expressions (7.63) and (7.70)–(7.71) in Section 7.4.2.

A natural follow-up question is whether there is a, perhaps a uniform, optimal estimator in the CRSS case. From straightforward algebra we find that

$$\mathrm{var}(\overline{\mu}) = \mu^2(1-c)^2\left(\frac{F}{1-F}\right)+\frac{\sigma^2}{n}\frac{1-2Fc+2Fc^2-F^2c^2}{2(1-F)}, \tag{7.72}$$

which is minimal for

$$c_{\text{opt}} = \frac{2\mu^2n+\sigma^2}{2\mu^2n+\sigma^2(2-F)}, \qquad d_{\text{opt}} = \frac{2\mu^2n+2\sigma^2}{2\mu^2n+\sigma^2(2-F)}. \tag{7.73}$$

In (7.72) and (7.73), $\sigma^2$ is the variance. It follows as either the first derivative of the mean function or, in the slightly more general case where there is an overdispersion parameter, as the first derivative of the mean multiplied with the overdispersion parameter.

Whereas constraint (7.67) on the pair $(c,d)$ does not depend on the particular exponential family considered, rather only on the constant probability of stopping, this is not true for the optimality condition (7.73). Because of its dependence on $\mu$ and $\sigma^2$, (7.73) will not generally allow for a uniform optimum, expect in specific examples. A few examples are given in Table 7.1. As Molenberghs *et al.* (2013) observed for

| Exp. fam. member | $c$ | $d$ |
|---|---|---|
| Normal | $\frac{2\mu^2 n + \sigma^2}{2\mu^2 n + \sigma^2(2-F)}$ | $\frac{2\mu^2 n + 2\sigma^2}{2\mu^2 n + \sigma^2(2-F)}$ |
| Bernoulli | $\frac{2\pi n + (1-\pi)}{2\pi n + (1-\pi)(2-F)}$ | $\frac{2\pi n + 2(1-\pi)}{2\pi n + (1-\pi)(2-F)}$ |
| Poisson | $\frac{2\lambda n + 1}{2\lambda n + 2 - F}$ | $\frac{2\lambda n + 2}{2\lambda n + 2 - F}$ |
| Exponential | $\frac{2n+1}{2n+2-F}$ | $\frac{2n+2}{2n+2-F}$ |

Table 7.1: *Coefficients for optimum unbiased generalized sample average estimators, in the case of a completely random sample size.*

the normal case, most solutions indeed indicate that there is no uniform minimum, even though all coefficients converge to 1 if the sample size increases. A noteworthy exception is the exponential family distribution, for which there is a uniform solution common to all values of the mean parameter and different from 1, for every value of the sample size $n$.

In all cases, when $F = 0$ then $d = 1$ and $c$ is irrelevant, while for $F = 1$, the reverse is true.

We have seen above that, even for CRSS, the sample average is not optimal, and that there is no uniform optimal solution, even though the sample average approximately is. The exponential case is an exception to this, as we saw above. However, the sample average is optimal in the restricted class of estimators that is invariant to future decisions. Indeed, if stopping occurs, then the choice of the coefficient $c$ leads to an unbiased estimator, provided the appropriate $d$ is chosen. However, this $d$ will never be used as it pertains to 'future' observations. This can be avoided only by setting both coefficients to be equal, from which the conventional sample average emerges.

The asymptotic behavior for a deterministic stopping rule is completely captured by the normal case, described in Section 7.4.2, because the stopping rule $F(k)$ has the effect of restricting the integrals over the stopping and continuation regions $\mathcal{S}$ and $\mathcal{C}$, respectively. This, together with the fact that $f_n(k)$ approaches a normal density with mean $n\mu$ and variance $n\sigma^2$ establishes this fact. As a result, we can restrict considerations regarding the deterministic case to the finite-sample situation. But also this one is very straightforward. Given that the joint distribution (7.14)–(7.15)

becomes (7.19)–(7.20), the functions $A_n(\mu)$ and $B_n(\mu)$ in (7.55) take the form:

$$A_n(\mu) = \int_{\mathcal{S}} f_n(k)dk, \qquad B_n(\mu) = \int_{\mathcal{S}} k f_n(k)dk. \tag{7.74}$$

and all results, such as marginal and conditional expectations of the GSA, carry over.

### 7.4.2 The Normal Case

Molenberghs *et al.* (2013) showed that expectation (7.53) of generalized sample average (7.51) becomes, for the normal case with probit stopping probability:

$$E(\overline{\mu}) = d\mu + (c-d)\mu\Phi(\nu) + \frac{2c-d}{2n}\frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) \tag{7.75}$$

with $\nu = (\alpha + \beta\mu)/\sqrt{1+\beta^2/n}$.

The specific case of a CRSS, here corresponding with $\beta = 0$, has been considered in Section 7.4.1.

When $\beta \neq 0$, expression (7.75) does not in general simplify. It is easy to see here that there cannot be a uniformly unbiased estimator, i.e., that there cannot exist $c$ and $d$ such that (7.75) reduces to $\mu$, for all $\mu$, and in particular for $\mu = 0$. For this special case

$$0 = \frac{2c-d}{2n} \cdot \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu_0),$$

where $\nu_0 = (\alpha)/\sqrt{1+\beta^2/n}$. Given that $\beta \neq 0$, this expression leads to the condition $2c = d$. Substituting this back into (7.75), which should be $\mu$ for every value of $\mu$, and not just for $\mu = 0$, produces $E(\overline{\mu}) = c\mu\,[2 - \Phi(\nu)]$, which equals $\mu$ only if $c = [2 - \Phi(\nu)]^{-2}$. Based on this, given that $\Phi(\nu)$ is not constant but rather depends on $\mu$, unless $\beta = 0$, we see that there can be no uniformly unbiased estimator for the generalized sample average type. In other words, a simple average estimator, that merely uses the observed measurements in a least-squares fashion, can never be unbiased unless $\beta = 0$.

Molenberghs *et al.* (2013) quantified the asymptotic bias. In Section 7.4.1 this was done in general for CRSS. Turning to the case of $\beta \neq 0$, Molenberghs *et al.* (2013) began with the ordinary sample average $c = d = 1$, which leads to expectation:

$$E(\widehat{\mu}) = \mu + \frac{1}{2n}\frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \phi(\nu) \overset{n\to+\infty}{\longrightarrow} \mu + \frac{1}{2n}\beta \cdot \phi(\alpha+\beta\mu) \overset{n\to+\infty}{\longrightarrow} \mu. \tag{7.76}$$

In particular, when $\beta \to +\infty$, we see that

$$E(\widehat{\mu}) = \mu + \frac{1}{2\sqrt{n}} \cdot \phi\left(\sqrt{n}\mu\right) \overset{n\to+\infty}{\longrightarrow} \mu. \tag{7.77}$$

There exist other choices that also lead to asymptotically unbiased generalized sample averages. For $\beta \neq 0$ but finite, the expectation becomes

$$E(\overline{\mu}) \overset{n\to+\infty}{\longrightarrow} d\mu + (c-d)\mu\Phi(\alpha+\beta\mu), \tag{7.78}$$

which equals $\mu$ if and only if:

$$d = \frac{1-c\Phi(\alpha+\beta\mu)}{1-\Phi(\alpha+\beta\mu)}. \tag{7.79}$$

While (7.79) and (7.67) are similar, there is a crucial difference between these: the latter is independent of $\mu$, while the former is not, except when $c = d = 1$. In other words, there is no uniformly asymptotically unbiased generalized sample average for finite, non-zero $\beta$, except for the ordinary sample average itself.

The above limits also follow from (7.62) and (7.63), because now $\eta(k/n) = \alpha + \beta k/n$ and the derivative therefore is $F'(n\mu) = \phi(\alpha+\beta\mu)\cdot\beta/n$, which leads to (7.76).

Molenberghs *et al.* (2013) also studied the deterministic stopping rule case, following from $\beta \to \infty$, because then (7.78) becomes

$$E(\overline{\mu}) = d\mu + (c-d)\mu\Phi(\sqrt{n}\mu) + \frac{2c-d}{2\sqrt{n}}\phi(\sqrt{n}\mu) \overset{n\to+\infty}{\longrightarrow} \begin{cases} c\mu & \text{if } \mu > 0, \\ d\mu & \text{if } \mu < 0, \\ 0 & \text{if } \mu = 0. \end{cases} \tag{7.80}$$

This provides us with the interesting situation that, for positive $\mu$, $c = 1$ yields an asymptotically unbiased estimator, regardless of $d$, with the reverse holding for negative $\mu$. In the special case that $\mu = 0$, both coefficients are immaterial. In addition, we see here as well that the only uniform solution is obtained by requiring that the bias asymptotically vanishes for all values of $\mu$, that is $c = d = 1$.

The pleasing asymptotic behavior of the sample average is connected to the choice of the stopping rule, in view of limiting expressions (7.63), (7.70), and (7.71). In this case, $\lim_{n\to+\infty} F(n\mu) = \Phi(\alpha+\beta\mu)$, a constant in $]0,1[$, while $\lim_{n\to+\infty} F'(n\mu) = (\beta/n)\phi(\alpha+\beta\mu) = 0$. Hence, the limits of $F'(n\mu)$, $F'(n\mu)/F(n\mu)$, and $F'(n\mu)/[1-F(n\mu)]$ are zero. The essence is that the stopping rule is a cumulative density function based transformation of a linear predictor in $k/n$. It is therefore of interes to examine the consequences of switching to a different class of stopping rule. Therefore, we change the stopping rule to $\Phi(\alpha+\beta k)$. Then $F'(n\mu) = \beta\phi(\alpha+\beta n\mu)$ which again tends to zero. However, depending on the sign of $\beta$ and $\mu$, $\Phi(\alpha+\beta n\mu)$ tends to either zero or one. Applying de l'Hôpital's rule to the case where $F(n\mu)$ tends to zero as well, produces $-\beta(\alpha+\beta n\mu)$ which tends to infinity, and hence the regularity condition

(7.70) appears *not* to be satisfied. This requires careful qualification, because not only does $F(n\mu)$ appear in (7.70), it is also the probability with which $N = n$, which then equally well tends to zero. Thus, for this case, in the limit, $E(\overline{\mu}|N = 2n) = E(\overline{\mu})$ and unbiasedness still applies. Evidently, when $1 - F(n\mu)$ tends to zero rather than $F(n\mu)$, we are in the mirror image of the above situation, and the result is the same. This result applies more generally. If $F(k) = \Phi(\alpha + \beta k n^m)$, with $m$ any real number, then $F'(n\mu)$ converges to zero whatever $m$ is. Further, $F(n\mu)$ converges to $\Phi(\alpha+\beta\mu)$ for $m = -1$, $\Phi(\alpha)$ for $m < -1$, and $\Phi(\pm\infty)$ (i.e., 0 or 1) for $m > -1$. This means that the sample average is asymptotically unbiased in all cases, and even conditionally asymptotically unbiased, based on the same logic as before.

### 7.4.3 The Binary Case

An explicit form for the expectation of the generalized sample average in the Bernoulli case is

$$E(\overline{\pi}) = d\pi + \frac{c}{n}\sum_{k=0}^{n} k \left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} F(k) - \frac{d}{2n}\sum_{k=0}^{2n} k\pi^k(1-\pi)^{2n-k} H(k), \quad (7.81)$$

with $H(k)$ as in (7.45).

The CRSS has been covered in Section 7.4.1, and the coefficients for optimal estimators listed in Table 7.1.

As an example, when stopping rule (7.13) is chosen, with $p = q = 0$ and $A(k) = k/n$, we have that $F(k) = A(k) = k/n$ and

$$H(k) = k/2n \cdot \left( \begin{array}{c} 2n \\ k \end{array} \right).$$

Hence, (7.81) becomes

$$\begin{aligned} E(\overline{\pi}) &= d\pi + \frac{c}{n^2}E_n(K^2) - \frac{d}{(2n)^2}E_{2n}(K^2) \\ &= \pi\left[d + \frac{c}{n}(1-\pi+n\pi) - \frac{d}{2n}(1-\pi+2n\pi)\right]. \end{aligned} \quad (7.82)$$

Clearly, the estimator is unbiased if and only if

$$d = \frac{1 - \frac{c}{n}(1-\pi+n\pi)}{1 - \frac{1}{2n}(1-\pi+2n\pi)}.$$

Hence, there is no uniform solution, neither in $\pi$ nor in $n$. When $n \to +\infty$,

$$d \overset{n\to+\infty}{\longrightarrow} \frac{1-c\pi}{1-\pi}. \quad (7.83)$$

Note that the ordinary sample average, i.e., $c = d = 1$, is a solution to (7.83), as it should.

Turning to the case of a deterministic stopping rule, assume that the stopping region $\mathcal{S}$ is defined by $(k \leq k_0)$, i.e., $F(k) = 1$ if $k \leq k_0$ and 0 otherwise. Functions $A_n(\pi)$ and $B_n(\pi)$ as in (7.74) are here:

$$A_n(\pi) = \sum_{k=0}^{k_0} \binom{n}{k} \pi^k (1-\pi)^{n-k} = \mathcal{I}(k_0, n, \pi), \tag{7.84}$$

$$B_n(\pi) = \sum_{k=0}^{k_0} k \binom{n}{k} \pi^k (1-\pi)^{n-k} = n\pi \mathcal{I}(k_0 - 1, n-1, \pi). \tag{7.85}$$

$\mathcal{I}(k, n, \pi)$, the binomial cumulative distribution function, is actually defined by (7.84). Various alternative formulations exist, but none is of direct use to us here. The expectation of the GSA becomes:

$$E(\overline{\pi}) = \pi \left[ d + \frac{2c-d}{2} \mathcal{I}(k_0 - 1, n-1, \pi) - \frac{d}{2} \mathcal{I}(k_0, n, \pi) \right]. \tag{7.86}$$

For the ordinary sample average, (7.86) reduces to

$$E(\overline{\pi}) = \pi \left\{ 1 + \frac{1}{2} \left[ \mathcal{I}(k_0 - 1, n-1, \pi) - \mathcal{I}(k_0, n, \pi) \right] \right\}.$$

## 7.5 Likelihood Estimators

### 7.5.1 The General Case

For notational convenience, we introduce the indicator variable $Z = I(N = n)$.

The joint likelihood for the observed data and stopping occurrence is:

$$L(\mu) = h_N(k) \exp\{\mu k - Na(\mu)\} \cdot F(k)^z \cdot [1 - F(k)]^{1-z}. \tag{7.87}$$

Likelihood decomposition (7.87) is of a selection model type. The factors pertaining to stopping are free of the mean parameter $\mu$. This simplifies the kernel of the likelihood $\ell(\mu)$, score function $S(\mu)$, and Hessian $H(\mu)$:

$$\ell(\mu) = \ln h_N(k) + \mu k - Na(\mu), \tag{7.88}$$

$$S(\mu) = k - Na'(\mu) = k - N\mu, \tag{7.89}$$

$$H(\mu) = -Na''(\mu) = -N\mu'. \tag{7.90}$$

The simplicity of this estimator is a direct consequence of ignorability. Based on (7.14)–(7.15), the conditional probability for the sample sum $K$, given the sample

size $N$, can be derived. For the case that $N = n$, the likelihood function is:

$$L_n(\mu) = \frac{F(k)h_n(k)e^{\mu k}}{\int F(k)h_n(k)e^{\mu k}dk}, \tag{7.91}$$

leading to the following expressions for the log-likelihood, score, and Hessian:

$$\begin{aligned}
\ell_n(\mu) &= \ln F(k) + \ln h_n(k) + \mu k - \ln \int F(k)h_n(k)e^{\mu k}dk, && (7.92)\\
S_n(\mu) &= k - \frac{B_n(\mu)}{A_n(\mu)} = k - E(K|N=n), && (7.93)\\
H_n(\mu) &= -\left[\frac{C_n(\mu)}{A_n(\mu)} - \left\{\frac{B_n(\mu)}{A_n(\mu)}\right\}^2\right] \\
&= -\left[E(K^2|N=n) - E(K|N=n)^2\right] = -\mathrm{var}(K|N=n). && (7.94)
\end{aligned}$$

Here $A_n(\mu)$ and $B_n(\mu)$ are as defined in (7.55), and

$$C_n(\mu) = \int k^2 f_n(k)F(k)dk.$$

When $N = 2n$, the likelihood takes the form:

$$L_{2n}(\mu) = \frac{D_n(\mu)}{1 - A_n(\mu)}, \tag{7.95}$$

with

$$D_n(\mu) = \exp\{\mu k - 2na(\mu)\}\left[h_{2n}(k) - \int h_n(z)h_n(k-z)F(z)dz\right].$$

Then, the counterparts to (7.92)–(7.94) are: Then, the counterparts to (7.92)–(7.94) are:

$$\begin{aligned}
\ell_{2n}(\mu) &= \mu k - 2na(\mu) + \ln\left[h_{2n}(k) - \int h_n(z)h_n(k-z)F(z)dz\right] \\
&\quad - \ln\left\{1 - A_n(\mu)\right\}, && (7.96)\\
S_{2n}(\mu) &= k - \frac{2n\mu - n\mu A_n(\mu) - B_n(\mu)}{1 - A_n(\mu)} \\
&= k - E(K|N=2n), && (7.97)\\
H_{2n}(\mu) &= -\frac{2n\sigma^2 - n\sigma^2 A_n(\mu) + (n\mu)^2 A_n(\mu) - C_n(\mu)}{1 - A_n(\mu)} \\
&\quad + \frac{[B_n(\mu) - n\mu A_n(\mu)][2n\mu - n\mu A_n(\mu) - B_n(\mu)]}{[1 - A_n(\mu)]^2} \\
&= -\left[E(K^2|N=2n) - E(K|N=2n)^2\right] = -\mathrm{var}(K|N=2n). && (7.98)
\end{aligned}$$

From the form of (7.93) and (7.97), it is immediately clear that the conditional expectations of the conditional scores are equal to zero and therefore also the marginal expectation.

The expectation of the joint likelihood based estimator, which is the ordinary sample average, was presented in Section 7.4.1. Even though there is small-sample bias in most cases different from CRSS, wide classes of stopping rules are asymptotically unbiased. The bias expressions in the conditional expectation of the sample average, which of course are also the bias expressions for the joint likelihood estimator, are of the form $E(K/N|N)-\mu$. These expressions coincide with the correction in conditional score equations (7.93) and (7.97) relative to (7.89), which follows immediately upon rewriting the former as $S_N(\mu) = k - N\mu + \{N\mu - E(K|N)\}$.

Turning to precision and information, first note that for CRSS, $H_n(\mu) = -n\sigma^2$ and $H_{2n}(\mu) = -2n\sigma^2$; hence the marginal and conditional information in this case reduces to $I(\mu) = I_c(\mu) = n\sigma^2(2 - F)$.

In the general case, the marginal and conditional information are

$$I(\mu) = n\sigma^2[2 - A_n(\mu)], \tag{7.99}$$

$$I_c(\mu) = n\sigma^2[2 - A_n(\mu)] - \frac{[n\mu A_n(\mu) - B_n(\mu)]^2}{A_n(\mu)[1 - A_n(\mu)]}. \tag{7.100}$$

Using information expressions (7.99)–(7.100), the bias for the marginal likelihood estimator, and the fact that the conditional likelihood estimator is unbiased, the mean squared error expressions are:

$$\mathrm{MSE}_n(\widehat{\mu}) = \frac{1}{n\sigma^2[2 - A_n(\mu)]} + \frac{1}{4n^2}\left[n\mu A_n(\mu) - B_n(\mu)\right]^2, \tag{7.101}$$

$$\mathrm{MSE}_n(\widehat{\mu}_c) = \frac{1}{n\sigma^2[2 - A_n(\mu)]} + \frac{[n\mu A_n(\mu) - B_n(\mu)]^2}{D_n(\mu)}, \tag{7.102}$$

where

$$\begin{aligned} D_n(\mu) = {} & A_n(\mu)[1 - A_n(\mu)]\{n\sigma^2[2 - A_n(\mu)]\}^2 \\ & -n\sigma^2[2 - A_n(\mu)][n\mu A_n(\mu) - B_n(\mu)]^2. \end{aligned}$$

Recall that for CRSS $B_n(\mu) = n\mu A_n(\mu)$ and both MSE expressions coincide. In the asymptotic case, (7.101)–(7.102) can be approximated, using (7.60)–(7.61), as:

$$\mathrm{MSE}_{n\to+\infty}(\widehat{\mu}) \simeq \frac{1}{n\sigma^2[2 - F(n\mu)]} + \frac{\sigma^4}{4}F'(n\mu)^2, \tag{7.103}$$

$$\mathrm{MSE}_{n\to+\infty}(\widehat{\mu}_c) \simeq \frac{1}{n\sigma^2[2 - F(n\mu)]} + \frac{F'(n\mu)^2}{E_n(\mu)}, \tag{7.104}$$

where, $E_n(\mu) = [2 - F(n\mu)]\left\{F(n\mu)[1 - F(n\mu)][2 - F(n\mu)] - n\sigma^2F'(n\mu)^2\right\}$.
Returning to the exact expressions (7.101)–(7.102), it is relatively straightforward to show that (7.101) is smaller than (7.102) if and only if $\sigma^2A_n(\mu)[1-A_n(\mu)][2-A_n(\mu)] \geq$

4. Requiring that this inequality is satisfied for all values of $A_n(\mu)$ in the unit interval comes down to requiring that $\sigma^2 \leq 2.54$. Hence, the MSE is smaller in the marginal case if the variance is sufficiently small. For binary data, for example this is always satisfied given that the variance takes the form $\pi(1-\pi)$. Also, asymptotically, $A_n(\mu)$ typically tends to either 0 or 1, and the above requirement is then also satisfied. In case $F'(n\mu)$ tends to zero as $n$ tends to infinity, both MSE expressions tend to the same limit.

### 7.5.2 The Normal Case

Molenberghs *et al.* (2013) studied this case in detail. Because of the relatively simple expressions for the normal density and the probit stopping rule (7.22), additional insight can be gained We summarize their arguments.

Joint-likelihood expressions (7.87)–(7.90) for this case are:

$$L(\mu) = \prod_{i=1}^{N} \phi(y_i;\mu) \cdot F\left(\alpha + \frac{\beta}{n}k\right)^z \cdot \left\{1 - F\left(\alpha + \frac{\beta}{n}k\right)\right\}^{1-z}, \quad (7.105)$$

$$\ell(\mu) = -\frac{1}{2}\sum_{i=1}^{N}(y_i - \mu)^2, \quad (7.106)$$

$$S(\mu) = \sum_{i=1}^{N}(y_i - \mu), \quad (7.107)$$

$$H(\mu) = -N. \quad (7.108)$$

By contrast, one can start from the conditional probability for the outcomes:

$$f(y_1,\ldots,y_N|Z=z) = \frac{\prod_{i=1}^{N}\phi(y_i;\mu)\Phi\left(\alpha+\frac{\beta}{n}k\right)^z\left[1-\Phi\left(\alpha+\frac{\beta}{n}k\right)\right]^{1-z}}{\Phi\left(\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}}\right)^z\left[1-\Phi\left(\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}}\right)\right]^{1-z}}. \quad (7.109)$$

Write, for convenience, $\widetilde{\alpha} = \alpha/\sqrt{1+\beta^2/n}$ and $\widetilde{\beta} = \beta/\sqrt{1+\beta^2/n}$. Further, let $\nu = \widetilde{\alpha} + \widetilde{\beta}\mu$. Consider first the case where $N = n$. The kernel of the likelihood $\ell(\mu)$,

score function $S(\mu)$, and Hessian $H(\mu)$ are:

$$\ell_n(\mu) = -\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^2 - \ln\Phi(\nu), \tag{7.110}$$

$$S_n(\mu) = \sum_{i=1}^{N}(y_i-\mu) - \widetilde{\beta}\cdot\frac{\phi(\nu)}{\Phi(\nu)}, \tag{7.111}$$

$$H_n(\mu) = -N + \widetilde{\beta}^2\cdot[\nu\cdot\Phi(\nu)+\phi(\nu)]\cdot\frac{\phi(\nu)}{\Phi(\nu)^2}. \tag{7.112}$$

When $N = 2n$, the corresponding expressions are:

$$\ell_{2n}(\mu) = -\frac{1}{2}\sum_{i=1}^{N}(y_i-\mu)^2 - \ln\left[1-\Phi(\nu)\right], \tag{7.113}$$

$$S_{2n}(\mu) = \sum_{i=1}^{N}(y_i-\mu) + \widetilde{\beta}\cdot\frac{\phi(\nu)}{1-\Phi(\nu)}, \tag{7.114}$$

$$H_{2n}(\mu) = -N - \widetilde{\beta}^2\cdot\{\nu\cdot[1-\Phi(\nu)]-\phi(\nu)\}\cdot\frac{\phi(\nu)}{\left[1-\Phi(\nu)\right]^2}. \tag{7.115}$$

Next, we consider bias, consistency, precision, and mean squared error of the joint and conditional likelihood estimators.

In the CRSS case, $\mu$ vanishes from the joint stopping model, and both estimators coincide with the ordinary sample average, amply studied in Section 7.4.

Asymptotic unbiasedness of the sample average follows simultaneously from direct calculation as well as from the fact that it is the maximum likelihood estimator from the joint likelihood (7.106). In terms of the conditional likelihood, the estimator is obtained from the solution to the score equations, (7.111) and (7.114). These can be reformulated as:

$$\widetilde{S}(\mu) = \frac{1}{N}\sum_{i=1}^{N}y_i - \mu - \frac{\beta}{\sqrt{1+\beta^2/n}}\cdot\phi(\nu)\cdot\left\{\frac{I(N=n)}{n\cdot\Phi(\nu)} - \frac{I(N=2n)}{2n\cdot[1-\Phi(\nu)]}\right\}. \tag{7.116}$$

The expectation of (7.116) results from (7.76), combined with the observation that the probability of stopping is $\Phi(\nu)$:

$$\begin{aligned} E[\widetilde{S}(\mu)] &= \mu + \frac{1}{2n}\frac{\beta}{\sqrt{1+\beta^2/n}}\cdot\phi(\nu) - \mu \\ &\quad -\frac{\beta}{\sqrt{1+\beta^2/n}}\cdot\phi(\nu)\cdot\left\{\frac{1}{n\cdot\Phi(\nu)}\cdot\Phi(\nu) - \frac{1}{2n\cdot[1-\Phi(\nu)]}\cdot[1-\Phi(\nu)]\right\} \\ &= 0. \end{aligned}$$

Finite-sample unbiasedness follows directly from the linearity of the score in the data. Thus, the difference between both score equations is bias-correcting. The correction

is a non-linear function of $\mu$ and has no closed-form solution, underscoring the point that no simple algebraic function of $K$ and $N$ will lead to the same estimator.

Finally, we note that the conditional likelihood estimator is also conditionally unbiased, i.e., it is unbiased for both situations $N = n$ and $N = 2n$ separately, in agreement with our results of Section 7.5.1. To see this explicitly, it is convenient to rewrite the expectation of the generalized sample average (7.75):

$$\begin{aligned} E(\overline{\mu}) &= c \cdot \left\{ \mu + \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)} \right\} \cdot \Phi(\nu) \\ &\quad + d \cdot \left\{ \mu - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{2n[-1\Phi(\nu)]} \right\} \cdot [1-\Phi(\nu)], \end{aligned}$$

from which both expectations $E(\overline{\mu}|N)$ follow:

$$E(\overline{\mu}|N=n) = \mu + \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{n\Phi(\nu)}, \tag{7.117}$$

$$E(\overline{\mu}|N=2n) = \mu - \frac{\beta}{\sqrt{1+\beta^2/n}} \cdot \frac{\phi(\nu)}{2n[1-\Phi(\nu)]}. \tag{7.118}$$

In conclusion, the sample average is conditionally and marginally biased, with the bias vanishing as $n$ goes to infinity, except in the situations that correspond to vanishing probabilities. In contrast, the conditional estimator is unbiased, whether considered conditionally on the observed sample size or marginalized over it.

Turning to precision, the expected information in the joint approach is

$$I(\mu) = E(N) = n \cdot \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu) + 2n \cdot [1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)] = n[2 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)], \tag{7.119}$$

where $\widetilde{\alpha}$ and $\widetilde{\beta}$ are as above. In the conditional case, this is

$$I_c(\mu) = n[2 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)] - \frac{\widetilde{\beta}^2 \phi(\widetilde{\alpha} + \widetilde{\beta}\mu)^2}{\Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)[1 - \Phi(\widetilde{\alpha} + \widetilde{\beta}\mu)]} \tag{7.120}$$

When $n \to \infty$, the information approaches

$$I_c(\mu) \overset{n\to+\infty}{\longrightarrow} n \left\{ [2 - \Phi(\alpha + \beta\mu)] - \frac{1}{n} \cdot \frac{\beta^2 \phi(\alpha + \beta\mu)^2}{\Phi(\alpha + \beta\mu)[1 - \Phi(\alpha + \beta\mu)]} \right\}.$$

The difference between joint and conditional information tends to zero when $n$ tends to infinity.

We conclude that the conditional estimator is less precise than the joint one, in contrast to many familiar settings such as contingency table analyses. The important feature here is that conditioning is done on a non-ancillary statistic. In line with the

general theory in Section 7.5.1, we have also seen that the joint approach leads to the ordinary sample average, an estimator that has met with considerable concern in the past in the sequential setting.

Because of the opposing results for bias and precision, it is useful to calculate the mean squared error for both estimators. The expressions from Molenberghs *et al.* (2013), are

$$\text{MSE}(\widehat{\mu}) = \frac{1}{n[2-\Phi(\nu)]} + \frac{1}{4n^2}\widetilde{\beta}^2\phi(\nu)^2, \tag{7.121}$$

$$\text{MSE}(\widehat{\mu}_c) \simeq \frac{1}{n[2-\Phi(\nu)]} + \frac{1}{[2-\Phi(\nu)]^2\Phi(\nu)[1-\Phi(\nu)]n^2}\widetilde{\beta}^2\phi(\nu)^2. \tag{7.122}$$

Comparing these, we see that $g(\nu) = [2-\Phi(\nu)]^2\Phi(\nu)[1-\Phi(\nu)] < 4$, the inequality being strict. In fact, the maximal value for $g(\nu)$ equals 0.619. Hence, the joint estimator has the smallest MSE of both, even though the difference will be very small for moderate to large sample sizes. This holds regardless of the choice for $\alpha$, $\beta$, and $n$, and of the true value of $\mu$. For $\beta$ finite and when $n \to \infty$, $\nu$ approaches $\alpha+\beta\mu$ and $\widetilde{\beta}$ approaches $\beta$. Then, $\Phi(\alpha+\beta\mu)$ and $\phi(\alpha+\beta\mu)$ become constant and the difference between the two expressions disappears because the second terms on the right hand sides of (7.121) and (7.122) are of the order of $1/n^2$.

To conclude this section, we examine the above quantities for the limiting case of a deterministic stopping rule, i.e., $\beta \to \pm\infty$. Focusing on the positive limit, we obtain

$$\frac{\alpha+\beta\mu}{\sqrt{1+\beta^2/n}} \overset{\beta\to+\infty}{\longrightarrow} \sqrt{n}\mu.$$

The marginal outcome model retains its normal-density form, while the other three expressions change. First, the conditional outcome model (7.109) becomes

$$f(y_1,\ldots,y_N|Z=z) = \frac{\prod_{i=1}^N \phi(y_i;\mu)}{\Phi\left(\sqrt{n}\mu\right)^z\left[1-\Phi\left(\sqrt{n}\mu\right)\right]^{1-z}}. \tag{7.123}$$

Second, (7.4) is given by $P(N=n|\boldsymbol{y}_i) = 1$ if $K > 0$ and 0 otherwise. Third, (7.22) takes the limiting form $P(N=n) = \Phi\left(\sqrt{n}\mu\right)$.

Molenberghs *et al.* (2013) derived the following bias expressions under deterministic stopping:

$$E(\overline{\mu}|N=n) = \mu + \frac{\phi(\sqrt{n}\mu)}{\sqrt{n}\Phi(\sqrt{n}\mu)}, \tag{7.124}$$

$$E(\overline{\mu}|N=2n) = \mu - \frac{\phi(\sqrt{n}\mu)}{2\sqrt{n}[1-\Phi(\sqrt{n}\mu)]}, \tag{7.125}$$

from which it follows that $E[\widetilde{S}(\mu)|N] = 0$. For the sample average a little more caution is required. From (7.117) and (7.118), it follows that $E(\overline{\mu}|N)$ converges to $\mu$ at a rate

of $n$, because $\nu \to \alpha+\beta\mu$. The situation is more subtle when $\beta \to \infty$. To show this, we take the limit of (7.124) and (7.125) as $n \to \infty$. When $\mu < 0$, applying de l'Hôpital's rule whenever needed, the limits are $E(\overline{\mu}|N = n) \to 0$ and $E(\overline{\mu}|N = 2n) \to \mu$. Similarly, when $\mu > 0$, the corresponding expressions are $E(\overline{\mu}|N = n) \to \mu$ and $E(\overline{\mu}|N = 2n) \to \mu/2$. It follows that when $\mu = 0$, these are both equal to 0. Somewhat surprising, this shows there is *no* bias in the conditional means: when $n \to \infty$, the probability itself that $N = n$ ($N = 2n$) for negative (positive) $\mu$ goes to zero. This implies that, overall, conditional inference based on the ordinary sample average is still acceptable.

For precision, the second term in (7.120) approaches

$$\frac{n\phi(\sqrt{n}\mu)^2}{\Phi(\sqrt{n}\mu)[1-\Phi(\sqrt{n}\mu)]}.$$

This term is non-zero for finite $n$ but can be shown to approach 0 if $n \to \infty$. This has the interesting consequence that there is no difference in precision when $\beta = 0$ and $\beta \to \infty$, but that there is for finite non-zero $\beta$.

For the mean squared error, the argument differs from the one used in the stochastic stopping rule case, because now $\nu = \sqrt{n}\mu$ and $\widetilde{\beta} = \sqrt{n}$, which leads to

$$\begin{aligned}
\mathrm{MSE}(\widehat{\mu}|\beta \to \infty) &= \frac{1}{n[2-\Phi(\sqrt{n}\mu)]} + \frac{1}{4n}\phi(\sqrt{n}\mu)^2, \\
\mathrm{MSE}(\widehat{\mu}_c|\beta \to \infty) &\simeq \frac{1}{n[2-\Phi(\sqrt{n}\mu)]} \\
&\quad + \frac{1}{[2-\Phi(\sqrt{n}\mu)]^2\Phi(\sqrt{n}\mu)[1-\Phi(\sqrt{n}\mu)]n}\phi(\sqrt{n}\mu)^2.
\end{aligned}$$

When $n \to \infty$, both expressions converge to $1/(2n)$ if the trial continues and $1/n$ if the trial stops, and the difference between them disappears.

### 7.5.3 The Binary Case

Joint-likelihood expressions for the binary case, in the probability parameter $\pi$ are:

$$L(\pi) = \binom{N}{k} \pi^k (1-\pi)^{N-k} \cdot F(k)^z \cdot [1-F(k)]^{1-z}, \tag{7.126}$$

$$\ell(\pi) \propto k \ln \pi + (N-k)\ln(1-\pi), \tag{7.127}$$

$$S(\pi) = \frac{k}{\pi} - \frac{N-k}{1-\pi}, \tag{7.128}$$

$$H(\pi) = \frac{-K + 2\pi K - \pi^2 N}{\pi^2(1-\pi)^2}. \tag{7.129}$$

The expected Hessian, for fixed sample size, is well known to be $-N/[\pi(1-\pi)]$. However, with our stopping rule $F(k) = k/n$, it can be shown to be

$$E[H(\pi)] = -\frac{(2-\pi)n}{\pi(1-\pi)}. \tag{7.130}$$

Likewise, given that the solution to $S(\pi)$ is the sample average, the bias is

$$\text{Bias}(\widehat{\pi}) = \frac{\pi(1-\pi)}{2n}. \tag{7.131}$$

We will return to this in what follows. Turning to the conditional expressions, for $N = n$, (7.91)–(7.94) become:

$$L_n(\pi) = \frac{\left( \begin{array}{c} n \\ k \end{array} \right) \pi^k (1-\pi)^{n-k} F(k)}{A_n(\pi)}, \tag{7.132}$$

with

$$A_n(\pi) = \sum_{\ell=0}^{n} \left( \begin{array}{c} n \\ \ell \end{array} \right) \pi^\ell (1-\pi)^{n-\ell} F(\ell),$$

leading to:

$$\ell_n(\pi) \propto k \ln \pi + (n-k)\ln(1-\pi) - \ln A_n(\pi), \tag{7.133}$$

$$S_n(\pi) = \frac{k}{\pi} - \frac{n-k}{1-\pi} - \frac{n\pi A_n(\pi) - B_n(\mu)}{A_n(\pi)}. \tag{7.134}$$

For the case where $N = 2n$ we obtain:

$$L_{2n}(\pi) = \frac{\left( \begin{array}{c} 2n \\ k \end{array} \right) \pi^k (1-\pi)^{2n-k} [1 - F(k)]}{1 - A_n(\pi)}, \tag{7.135}$$

$$\ell_{2n}(\pi) \propto k \ln \pi + (2n-k)\ln(1-\pi) - \ln[1 - A_n(\pi)], \tag{7.136}$$

$$S_{2n}(\pi) = \frac{k}{\pi} - \frac{2n-k}{1-\pi} + \frac{n\pi A_n(\pi) - B_n(\mu)}{1 - A_n(\pi)}. \tag{7.137}$$

The fact that $E[S_N(\pi)|N] = 0$ follows from the derivations in Section 7.5.1, as well as from first principles.

It is clear that the above expressions are slightly different than the general expressions (7.87)–(7.90), because $\pi$ is not the natural parameter. This does not prohibit further derivations but makes them cumbersome from an algebraic standpoint. Therefore, we switch to the logit form, i.e., $\alpha = \ln[\pi/(1-\pi)]$ will be used. Furthermore, we restrict attention to the particular stopping rule used in previous sections,

$F(k) = k/n$. Then, (7.126)–(7.129) become:

$$L(\alpha) = \left( \begin{array}{c} N \\ k \end{array} \right) \frac{e^{\alpha k}}{(1+e^{\alpha})^N} \cdot \left(\frac{k}{n}\right)^z \cdot \left(\frac{n-k}{n}\right)^{1-z}, \tag{7.138}$$

$$\ell(\alpha) \propto \alpha k - N \ln(1+e^{\alpha}), \tag{7.139}$$

$$S(\alpha) = k - N\pi, \tag{7.140}$$

$$H(\alpha) = -N\pi(1-\pi). \tag{7.141}$$

The use of $\pi$ on the right hand sides of (7.140) and (7.141) rather than $\alpha$ is for convenience only. The expected Hessian is straightforward to derive, given that $E(N) = n(2-\pi)$:

$$E[H(\alpha)] = -n(2-\pi)\pi(1-\pi). \tag{7.142}$$

In fact, this calculation is considerably easier than the derivation of (7.130), even though they are equivalent. Indeed, (7.130) follows from (7.142) by applying the delta method. Because $\pi = \text{expit}(\alpha)$, the derivative is $\Delta = \partial\alpha/\partial\pi = [\pi(1-\pi)]^{-1}$, and $E[H(\pi)] = \Delta^2 \cdot E[H(\alpha)]$, as it should.

The forms for (7.133)–(7.137), supplemented with the Hessians, are:

$$\ell_n(\alpha) \propto \alpha k - n\ln(1+e^{\alpha}) - \alpha + \ln(1+e^{\alpha}), \tag{7.143}$$

$$S_n(\alpha) = (k-1) - (n-1)\pi, \tag{7.144}$$

$$H_n(\alpha) = -(n-1)\pi(1-\pi), \tag{7.145}$$

$$\ell_{2n}(\alpha) \propto \alpha k - 2n\ln(1+e^{\alpha}) + \ln(1+e^{\alpha}), \tag{7.146}$$

$$S_{2n}(\alpha) = k - (2n-1)\pi, \tag{7.147}$$

$$H_{2n}(\alpha) = -(2n-1)\pi(1-\pi). \tag{7.148}$$

Note that the conditional Hessians are in line with what one would expect from conditioning upon the sample size: one 'degree of freedom' is removed for mean parameter estimation. Such an operation though, is standard only when the sample size is fixed. The counterintuitive effect on the efficiency was seen in general in Section 7.5.1 and very explicitly for the normal data setting in Section 7.5.2. Straightforward algebra then establishes:

$$E[H_N(\alpha)] = -\pi(1-\pi)[(2-\pi)n - 1] = E[H(\alpha)] + \pi(1-\pi). \tag{7.149}$$

Thus, the conditional information is expected to take one subject less into account than the marginal expectation, precisely the opposite of what one would expect in the fixed sample-size case. The bias in the estimators is easy to quantify, given that

the estimators are $\widehat{\pi} = k/N$ in the marginal case and $\widehat{\pi}_c = (k-1)/(n-1)$ when $N = n$ and $\widehat{\pi}_c = k/(2n-1)$ when $N = 2n$. The biases are $(n-k)/([n(n-1)]$ and $-k/[2n(2n-1)]$, respectively. This follows from the difference between the marginal and conditional estimators, given that the latter is unbiased. For this stopping rule, $E(K|N=n) = n\pi + 1 - \pi$ and $E(K|N=2n) = \pi(2n-1)$, and so the average bias is (7.131), as we expect.

The variances are equal to the negative inverses of the expected Hessians. These, combined with bias (7.131), readily leads to the MSE. Of course, (7.145) and (7.148) are for $\widehat{\alpha}_c$ and hence the delta method needs to be applied to obtain the variances for $\widehat{\pi}_c$. Note that the variance for $\widehat{\pi}$ was already derived in (7.130), but applying the delta method to (7.141) gives the exact same result. The additional expressions are

$$\text{var}(\widehat{\pi}_c|N=n) = \frac{\pi(1-\pi)}{n-1}, \tag{7.150}$$

$$\text{var}(\widehat{\pi}_c|N=2n) = \frac{\pi(1-\pi)}{2n-1}, \tag{7.151}$$

with the expected conditional Hessians the inverses of these quantities:

$$\begin{aligned} \text{var}(\widehat{\pi}_c) &= \text{var}(\widehat{\pi}_c|N=n)\cdot\pi + \\ &\quad \text{var}(\widehat{\pi}_c|N=2n)\cdot(1-\pi) \\ &= \frac{(n\pi+n-1)}{(n-1)(2n-1)}\pi(1-\pi). \end{aligned} \tag{7.152}$$

Note that the derivation of overall variance (7.153) involves the expectation of the conditional variances only, while the variance of the conditional expectations is zero, because both conditional estimators are unbiased. Finally,

$$\begin{aligned} \text{MSE}(\widehat{\pi}) &= \text{var}(\widehat{\pi}) + \text{Bias}(\widehat{\pi}) \\ &= \frac{\pi(1-\pi)}{n(2-\pi))} + \frac{\pi^2(1-\pi)^2}{(2n)^2} = \frac{1}{n}\cdot\frac{\pi(1-\pi)}{2-\pi} + \mathcal{O}(n^{-2}), \end{aligned} \tag{7.153}$$

$$\begin{aligned} \text{MSE}(\widehat{\pi}_c) &= \text{var}(\widehat{\pi}_c) = \frac{(n\pi+n-1)\pi(1-\pi)}{(n-1)(2n-1)} \\ &= \frac{1}{n}\cdot\frac{\pi(1-\pi)(\pi+1)}{2} + \mathcal{O}(n^{-2}). \end{aligned} \tag{7.154}$$

Calculating the difference between (7.154) and (7.153), we find

$$\text{MSE}(\widehat{\pi}_c) - \text{MSE}(\widehat{\pi}) = \frac{1}{n}\frac{\pi^2(1-\pi)^2}{2(2-\pi)} + \mathcal{O}(n^{-2}).$$

Hence, like in the normal case, the joint estimator is more efficient than the marginal one. Of course, the MSE increase when moving from the joint to the conditional

estimator is modest, with $\text{MSE}(\widehat{\pi}) \leq \text{MSE}(\widehat{\pi}_c) \leq 1.125\text{MSE}(\widehat{\pi})$, the maximum discrepancy reached for $\pi = 0.5$, and equality for $\pi = 0$ or $\pi = 1$. Because the stopping rule depends on $K_n$, and because (7.4) combined with the conditional outcome model is a pattern-mixture factorization (7.2), $N$ is not ancillary to $K$.

## 7.6 Discussion

we have considered the consequences for statistical inference of a random sample size. Our setting is that of univariate random variables from the exponential family that are subject to a stopping rule such that the sample size is either $N = n$ or $N = 2n$, with $n$ specified by design. Though the context considered here is for a single group trial, $n$ can specified by other designs, such as two-arm clinical trials, whether parallel or cross over. The stopping rule is stochastic and is allowed to depend on the sample sum $K$ over the first $n$ observations. The rule is generic in the sense that its limiting cases are a deterministic stopping rule, such as in a sequential trial, and a completely random sample size, independent of the data. This setting extends those of both Liu *et al.* (2006) and Molenberghs *et al.* (2013); the former restrict attention to a deterministic stopping rule, although they do so for an arbitrary number of interim looks. The latter confined attention to normally distributed outcomes only.

We have focused on three important inferential aspects. First, we have shown that the sufficient statistic $(K, N)$ is incomplete. Second, we have examined the consequences of this for the sample average, as well as for linear generalizations thereof. We have shown that there is small-sample bias, except for the CRSS case. Even then, there is no optimal estimator, except for the exponential distribution, for which the optimum differs from the ordinary sample average. Third, we have studied maximum likelihood estimation in both a joint as well as a conditional framework. The joint likelihood is for the exponential-family parameter and the stopping rule simultaneously. The conditional likelihood starts from the conditional distribution of the outcomes, given the sample size. Also here, counterintuitive results are derived. The joint likelihood produces the sample average as maximum likelihood estimator, which is biased in finite samples but is asymptotically unbiased, provided a regularity condition on the stopping rule applies. The conditional likelihood estimator is unbiased, even in small samples. This notwithstanding, the sample average has smaller MSE than the conditional estimator in many important cases, such as the normal and binary examples considered, as well as when the variance of the outcomes is sufficiently small. Under regularity conditions, both estimators are asymptotically equivalent,

with the difference between both being $\mathcal{O}(n^{-1})$. The regularity condition is not very restrictive; it essentially comes down to requiring that $F'(k = n\mu)$ approaches zero where $F$ is the stopping rule. For broad classes of parametric functions, this condition is satisfied. We have shown that the corresponding conditional expectations are unbiased.

Hence, when the regularity conditions are satisfied, the sample average remains an attractive and sensible choice for sequential trials. Thus, while some familiar inferential properties no longer hold, estimation after sequential trials is more straightforward than commonly considered and there is little need for complicated, modified estimators, given that the ordinary sample average is acceptable for wide classes of stopping rules, whether stochastic or deterministic.

Molenberghs *et al.* (2013) considered several ramifications of their developments. They commented on the situation of an arbitrary number of looks in a sequential trial, and considered in detail the CRSS case for more than two possible sample sizes. All of this was done for normally distributed outcomes. They also commented on the connection between their derivations and longitudinal outcomes subject to dropout of an MAR type, where dropout depends on observed but not further on unobserved outcomes. While similar, there are subtle differences because now the randomness in the sample size pertains to the number of measurements per subject, rather than to the number of subjects. The difference lies in the fact that measurements within a subject are not independent. Our results extend to these settings as well for the exponential family. Furthermore, connections can be made with a variety of other settings with random sample sizes, such as clustered data with informative cluster sizes, time-to-event data subject to censoring, jointly observed longitudinal and time-to-event data, and random observation times. These settings are currently scrutinized further, and will be reported in a separate manuscript

# Chapter 8

# Estimation After a Group Sequential Trial

Group sequential trials(GST) are the most popular form of RSS trials in medical research and estimation after such trial is still hot topic research area. Using the findings in Chapter 7, we shall in this chapter, that estimation after GST can be as flexible as after a FSS trial.

## 8.1 Introduction

Principally for ethical and economic reasons, group sequential clinical trials are in common use (Wald, 1945; Armitage, 1975; Whitehead, 1997; Jennison and Turnbull, 2000). Tools for constructing such designs, and for testing hypotheses from the resulting data, are well established both in terms of theory and implementation. By contrast, issues still surround the problem of estimation (Siegmund, 1978; Hughes and Pocock, 1988; Todd, Whitehead, and Facey, 1996; Whitehead, 1999) following such trials. In particular, various authors have reported that standard estimators such as the sample average are biased. In response to this, various proposals have been made to remove or at least alleviate this bias and its consequences (Tsiatis, Rosner, and Mehta, 1984; Rosner and Tsiatis, 1988; Emerson and Fleming, 1990). An early suggestion was to use a conditional estimator for this Blackwell (1947).

To successfully address the bias issue, it is helpful to understand its origins. Lehman (1950) showed that it stems from the so-called *incompleteness* of the sufficient statistics involved, which in turn implies that there can be no minimum vari-

ance unbiased linear estimator. Liu and Hall (1999) and Liu *et al.* (2006) explored this incompleteness in group sequential trials, for outcomes with both normal and one-parameter exponential family distributions. For these distributions, Chapter7 and Molenberghs *et al.* (2013) and embedded the problem in the broader class with random sample size, which includes, in addition to sequential trials, incomplete data, completely random sample sizes, censored time-to-event data, and random cluster sizes. In so doing, they were able to link incompleteness to the related concepts of ancillarity and ignorability in the missing-data sense. By considering the conventional sequential trial with a deterministic stopping rule as a limiting case of a stochastic stopping rule, these authors were able to derive properties of families of linear estimators as well as likelihood-based estimators. The key results are as follows: (1) the sample average (SA) exhibits finite sample bias, although it is asymptotically unbiased; (2) apart from the exponential distribution setting, there is no optimal linear estimator, although the sample average is asymptotically optimal; (3) the validity of the sample average as an estimator also follows from standard ignorable likelihood theory; (4) there exists a maximum likelihood estimator that conditions on the realized sample size (CL), which is finite sample unbiased, but has slightly larger variance and mean square error (MSE) than the SA.

There is a subtle issue surrounding the properties of the SA. Evidently, the CL is unbiased both conditionally and marginally with respect to the sample size. By contrast, the SA is marginally unbiased, but there exist classes of stopping rules where, conditionally on the sample size, there is asymptotic bias for some values of the sample sizes. Surprisingly, this is not of concern. In Chapter 7, we showed this for the case of two possible sample sizes, $N = n$ and $N = 2n$. With such a stopping rule, it is possible that, for example when $N = n$, the bias grows unboundedly with $n$; when this happens though, the probability that $N = n$ shrinks to 0 at the same rate. As a consequence, in large but finite samples, simulations have confirmed this bias, which has led to attempts to make corrections. The developments in Molenberghs *et al.* (2013) and those in Chapter 7, however, show that a correction is not strictly necessary since SA, because of its likelihood basis, it can be used in conjunction with standard likelihood-based measures of precision, such as standard errors and associated confidence intervals to provide valid inferences. If, on the other hand, strict finite sample unbiasedness is regarded as essential, the conditional MLE can be used, which, like MLE, also admits the standard likelihood-based precision measures. This is a very important result and should be contrasted with the various *ad hoc* precision estimators that have been developed in the past. Although there is a mild computation cost involved, being a likelihood estimator, CL follows from a general

principle and avoids the *ad hoc* nature of existing alternatives, which themselves are no simpler computationally, and some are considerably more demanding. Thus, in spite of some less conventional properties, the SA and its associated precision measures can be used **in some settings** without problem.

A major limitation of Molenberghs *et al.* (2013) and results in Chapter 7 is the restriction to two looks of equal size. It is the main aim of this paper to extend this work to the practically more useful setting of multiple looks of potentially different sample sizes.

## 8.2 Problem and Model Formulation

Consider a sequential trial with $L$ pre-specified looks, with sample sizes $n_1 < n_2 < \ldots, < n_L$. Assume that there are $n_j$ i.i.d. observations $Y_1, \ldots, Y_{n_j}$, from the $j$th look that follow an exponential family distribution with density

$$f_\mu(y) = h(y) \exp\{\mu y - a(\mu)\}, \tag{8.1}$$

for $\mu$ the natural parameter, $a(\mu)$ the mean generating function, and $h(y)$ the normalizing constant.

Subsequent developments are based on a generic data-dependent stochastic stopping rule, which we write

$$\pi(N = n_j | k_{n_j}) = F\left(k_{n_j} \middle| \boldsymbol{\psi}\right) = F\left(k_{n_j}\right), \tag{8.2}$$

where $K_{n_j} = \sum_{i=1}^{n_j} Y_i$ also has an exponential family density:

$$f_{n_j}(k) = h_{n_j}(k) \exp\left\{\mu k_{n_j} - n_j a(\mu)\right\}. \tag{8.3}$$

Our inferential target is the parameter $\mu$, or a function of this.

### 8.2.1 Stochastic Rule As A Group Sequential Stopping Rule

. While we do not need to provide an explicit expression for the stopping rule at this point, as our developments apply to a broad class, it is useful to note that In Chapter 7, have studied in detail the behavior of stopping rules taking the form $F\left(\alpha_j + \beta k_{n_j}/n_j^m\right)$, for some power $m$ where it was shown that the limiting cases, $\beta \to \infty$ and $\beta \to -\infty$ correspond to deterministic stopping rules commonly used in sequential trials. In the specific example of normally distributed responses, we assume $F\left(k_{n_j} \middle| \boldsymbol{\psi}\right) = \Phi\left(\alpha_j + \beta k_{n_j}/n_j\right)$, where $m = 1$, and $\boldsymbol{\psi} = (\alpha, \beta)$ is specified by design

and $\Phi(.)$ represent the cumulative probability function for the normal distribution. Clearly, if $\beta$ is too large, the probability of stopping is either 0 or 1 depending on the value of $k_{n_j}/n_j$, which is similar to group sequential trial rules, where one decides to stop or continue based on the observed statistic and pre-specified boundaries. The value of $\alpha$ is paramount to deciding the behavior of stopping boundaries. Consider O'Brien and Fleming stopping boundaries where it is difficult to stop in early stages; one can then specify $\alpha_j$ such that the probability of stopping increases with the stages. In addition to the computational advantages and the associated practicality, we use the stochastic rule to maintain the focus of this paper, which is estimation.

## 8.3 Incomplete Sufficient Statistics

Several concepts play a crucial role in determining the properties of estimators following sequential trial: incompleteness, a missing at random (MAR) mechanism, ignorability, and ancillarity (Molenberghs *et al.*, 2013). We consider the role of incompleteness first: a statistic $s(Y)$ of a random variable $Y$, with $Y$ belonging to a family $P_\mu$, is complete if, for every measurable function $g(\cdot)$, $E[g\{s(Y)\}] = 0$ for all $\mu$, implies that $P_\mu[g\{s(Y)\} = 0] = 1$ for all $\mu$ (Casella and Berger, 2001, pp. 285–286). Incompleteness is central to the various developments (Liu and Hall, 1999; Liu *et al.*, 2006; Molenberghs *et al.*, 2013) because of the the Lehman-Scheffé theorem which states that "if a statistic is unbiased, complete, and sufficient for some parameter $\mu$, then it is the best mean-unbiased estimator for $\mu$," (Casella and Berger, 2001). In the present setting, the relevant sufficient statistic is not complete, and so the theorem can *not* be applied here.

In line with extending the work of Molenberghs *et al.* (2013) and findings in Chapter 7, to a general number of looks, we explore incompleteness and its consequencies in studies with more than two looks using the stochastic rule.
In a sequential setting, a convenient sufficient statistic is $(K, N)$. Following the developments in the above papers, the joint distribution for $(K, N)$ is:

$$p(K,N) = f_0(K,N)\, F(K_N) \tag{8.4}$$

$$f_0(k_{n_1}, n_1) = f_{n_1}(k_{n_1}) \tag{8.5}$$

$$f_0(k_{n_j}, n_j) = \int f_0(k_{n_{j-1}}, n_{j-1}) f_{n_j - n_{j-1}}(k_{n_j} - k_{n_{j-1}}) \left[1 - F(k_{n_{j-1}})\right] dk_{n_{j-1}}. \tag{8.6}$$

If $(K, N)$ were complete, then there would exist a function $g(K, N)$ such that

$E\left[g(K,N)\right]=0$ if and only if $g(K,N)=0$, implying that

$$\begin{aligned} 0 &= \int g(k_{n_1},n_1)f_{n_1}(k_{n_1})F(k_{n_1})dk_{n_1}+\sum_{j=2}^{L-2}\int g(k_{n_j},n_j)H(k_{n_j})F(k_{n_j})dk_{n_j} \\ &\quad +\int g(k_{n_L},n_L)H(k_{n_L})F(k_{n_L})dk_{n_L}, \end{aligned} \tag{8.7}$$

with

$$H(k_{n_j})= \left[\int\underbrace{\ldots}_{j-1}\int f_0(k_{n_{j-1}},n_{j-1})f_{n_j-n_{j-1}}(k_{n_j}-k_{n_{j-1}})\left[1-F(k_{n_{j-1}})\right]dk_{n_1}\ldots dk_{n_{j-1}}\right].$$

Substituting the general exponential form (8.3) into (8.7), and applying properties of exponential family probability distribution, gives

$$\begin{aligned} 0 &= \int h_{n_L-n_1}e^{(\mu k_{n_1})}\int g(k_{n_1},n_1)F(k_{n_1})h_{n_1}(k_{n_1})\exp(\mu k_{n_1})dk_{n_1} \\ &\quad +\sum_{j=2}^{L-2}\int h_{n_L-n_j}e^{(\mu k_{n_j})}\int g(k_{n_j},n_j)\widetilde{H}(k_{n_j})\exp(\mu k_{n_j}-n_j)F(k_{n_j})dk_{n_j} \\ &\quad +\int g(k_{n_L},n_L)\widetilde{H}(k_{n_L})\exp(\mu k_{n_L})F(k_{n_L})dk_{n_L}, \end{aligned} \tag{8.8}$$

where

$$\widetilde{H}(k_{n_j})=\left[\int\underbrace{\ldots}_{j-1}\int\prod_{i=1}^{j-1}h_{n_1}(k_{n_1})h_{n_{i+1}-n_i}(k_{n_{i+1}}-k_{n_i})\left[1-F(k_{n_i})\right]dk_{n_1}\ldots dk_{n_{j-1}}\right].$$

The right hand side is a convolution and making use of properties of linearity and uniqueness of the Laplace transform it can be shown that:

$$\begin{aligned} g(k_{n_L},n_L)\widetilde{H}(k_{n_L}) &= -\sum_{j=1}^{L-1}\int g(z_j,n_j)\widetilde{H}(z_j)F(z_j)dz_j, \\ g(k_{n_L},n_L) &= \frac{\sum_{j=1}^{L-1}\int g(z_j,n_j)\widetilde{H}(z_j)F(z_j)dz_j}{\widetilde{H}(k_{n_L})}. \end{aligned}$$

Assigning, for example, arbitrary constants to $g(n_1,k_{n_1})\ldots g(n_{L-1},k_{n_{L-1}})$, a value can be found for $g(n_L,k_{n_L})\neq 0$, contradicting the requirement for $(K,N)$ to be complete, hence establishing incompleteness. From applying the Lehmann-Scheffé theorem, that no best mean-unbiased estimator is guaranteed to exist. The practical consequence of this is that even estimators as simple as a sample average need careful consideration and comparison with alternatives. For this, we embed the sample average in a broader class of linear estimator, and also study it from a likelihood perspective.

## 8.4 Generalized Sample Averages

Extending the definition in Molenberghs *et al.* (2013), the generalized sample average (GSA) can be be defined as:

$$\bar{\mu}_g = \sum_{j=1}^{L} \frac{a_j}{n_j} k_{n_j}, \tag{8.9}$$

for a set of constants $a_1, \ldots, a_L$. The SA follows as the special case where each $a_j = 1$. To explore the properties of the GSA we make use of the fact that:

$$\int f_{n_1}(k_{n_1}) dk_{n_1} + \sum_{j=2}^{L} \int \widetilde{H}(k_{n_j}) F(k_{n_j}) \exp(\mu k_{n_j} - n_j a(\mu)) dk_{n_j} = 1,$$

and derive three useful identities:

$$\int f_{n_1}(k_{n_1}) dk_{n_1} = 1 - \sum_{j=2}^{L} A_{n_j}(\mu), \tag{8.10}$$

$$\sum_{j=1}^{L} B_{n_j}(\mu) = \sum_{j=1}^{L} n_j a'(\mu) A_{n_j}(\mu), \tag{8.11}$$

$$\sum_{j=1}^{L} C_{n_j} = \sum_{j=1}^{L} 2 n_j a'(\mu) B_{n_j}(\mu) - [n_j a'(\mu)]^2 A_{n_j}(\mu) + n_j a''(\mu) A_{n_j}(\mu),$$

where

$$\begin{aligned}
A_{n_1}(\mu) &= \int f_{n_1}(k_{n_1}) dk_{n_1}, \\
B_{n_1}(\mu) &= \int k_{n_1} f_{n_1}(k_{n_1}) dk_{n_1}, \\
C_{n_1}(\mu) &= \int k_{n_1}^2 f_{n_1}(k_{n_1}) dk_{n_1}, \\
A_{n_j}(\mu) &= \int \widetilde{H}(k_{n_j}) F(k_{n_j}) \exp(\mu k_{n_j} - n_j a(\mu)) dk_{n_j}, \quad (j > 1) \\
B_{n_j}(\mu) &= \int k_{n_j} \widetilde{H}(k_{n_j}) F(k_{n_j}) \exp(\mu k_{n_j} - n_j a(\mu)) dk_{n_j}, \\
C_{n_j}(\mu) &= \int k_{n_j}^2 \widetilde{H}(k_{n_j}) F(k_{n_j}) \exp(\mu k_{n_j} - n_j a(\mu)) dk_{n_j} \quad (j > 1).
\end{aligned}$$

Using identities (8.10) and (8.11), the expectation of (8.9) can then be formulated as

$$\begin{aligned}
E[\bar{\mu}_g] &= \frac{a_1}{n_1} B_{n_1}(\mu) + \sum_{j=2}^{L} \frac{a_j}{n_j} B_{n_j}(\mu) \qquad (8.12) \\
&= a_1 \mu + \sum_{j=2}^{L} a_1 A_{n_j}(\mu) \frac{n_1 - n_j}{n_1} \left[ \frac{n_1 a_j - n_j a_1}{a_1 (n_1 - n_j)} E\left\{ \frac{K}{N} \middle| N = n_j \right\} - \mu \right],
\end{aligned}$$

establishing the bias as a function of the difference between the marginal and conditional means. When (8.12) is unbiased, at least one value among $a_1, \ldots, a_L$ will depend on $\mu$. This means that none of the GSA can be uniformly unbiased. Focusing on the SA, the expectation reduces to

$$\begin{aligned} E\left[\bar{\mu}\right] &= \mu + \sum_{j=2}^{L} A_{n_j}(\mu) \frac{n_1 - n_j}{n_1} \left[ \frac{B_{n_j}(\mu)}{n_j A n_j(\mu)} - \mu \right] \\ &= \mu + \sum_{j=2}^{L} A_{n_j}(\mu) \frac{n_1 - n_j}{n_1} \left[ E\left\{ \frac{K}{N} \middle| N = n_j \right\} - \mu \right], \end{aligned} \tag{8.13}$$

from which we get the bias as

$$\begin{aligned} \sum_{j=2}^{L} A_{n_j}(\mu) \frac{n_1 - n_j}{n_1} \left[ \frac{B_{n_j}(\mu)}{n_j A n_j(\mu)} - \mu \right] &= \sum_{j=2}^{L} \frac{n_1 - n_j}{n_1 n_j} \left[ B_{n_j}(\mu) - A_{n_j} n_j \mu \right] \\ &= \sum_{j=1}^{L} \frac{\left[ B_{n_j}(\mu) - A_{n_j} n_j \mu \right]}{n_j}. \end{aligned} \tag{8.14}$$

Thus, the SA is unbiased when the conditional and marginal means are equal.

## 8.5 Likelihood Estimation

We now consider the marginal, or joint maximum likelihood estimator, as well as its conditional counterpart. Likelihood methods, while allowing for a unified treatment across a variety of settings (e.g., data types, stopping rules), they do rely heavily on correct parametric specification. This should be taken into account when opting for a particular approach.

Our results consider the MLE from a specific angle, that of ignorability, and take a deterministic stopping rule as a limiting case. This provides additional information in a setting where some take the MLE for granted, while others prefer different, often tailor made estimators.

Molenberghs *et al.* (2013) connected sequential trials and incomplete data theory and a brief summary was presented in Section **??**. Using a selection model factorization for the joint distribution of observed data and sample size, they showed that separability and ignorability hold, such that under a missing at random (MAR) assumption, maximizing the joint likelihood is equivalent to maximizing the likelihood of the observed data only. This connection is crucial when considered against the background of Kenward and Molenberghs (1998), where it was shown that under frequentist inference and the missing at random (MAR) assumption, the observed

information matrix gives valid inferences. Provided that use is made of the likelihood ratio, Wald or score statistics based on the observed information, then reference to a null asymptotic $\chi^2$ distribution will be appropriate because this is derived from the implicit use of the unconditional sampling framework. We therefore explore joint modeling of observed data and sample sample size for a general number of looks, the properties of which will be compared to the likelihood of observed data conditional on the sample size.

### 8.5.1 Joint Likelihood

The joint distribution of the sufficient statistics $(K, N)$ is given by;

$$f(K,N) = h_N(K)\exp\left[K\mu - Na(\mu)\right] \cdot \prod_{i=1}^{L-1}\left[1 - F(k_{n_j})\right]F(k_{n_L})^{I(i<L)}. \tag{8.15}$$

Because our stopping rule is independent of the parameter of interest, the log-likelihood, the score, the Hessian, and the expected information simplify as follows:

$$\begin{aligned}
\ell(\mu) &= \ln[h_N(K)] + K\mu - Na(\mu) && (8.16)\\
&+ \ln\left\{\prod_{i=1}^{L-1}\left[1 - F(k_{n_j})\right]F(k_{n_L})^{I(i<L)}\right\},\\
S(\mu) &= K - Na'(\mu), && (8.17)\\
H(\mu) &= -Na''(\mu), && (8.18)\\
I(\mu) &= \sum_{j=1}^{L} n_j a''(\mu)A_{n_j}(\mu). && (8.19)
\end{aligned}$$

In deriving the score (8.17) from (8.17) the rightmost term drops out, *i.e.*, conventional ignorability applies. As a consequence, the maximum likelihood estimator reduces to $\widehat{\mu} = a'(\mu) = K/N$, the SA. Under the usual likelihood regularity conditions, the SA is then consistent and asymptotically normally distributed, and the likelihood-based precision estimator and its corresponding confidence interval are valid. In other words, this conventional asymptotic behavior contrasts with the idiosyncratic small-sample properties of the GSA derived in the previous section.

Because of the bias, a finite sample comparison among estimators needs to be based on the MSE. For $\widehat{\mu}$, this is

$$\mathrm{MSE}(\widehat{\mu}) = \frac{1}{\sum_{j=1}^{L} n_j a''(\mu)A_{n_j}(\mu)} + \left[\sum_{j=1}^{L}\frac{\left[B_{n_j}(\mu) - A_{n_j}n_j\mu\right]}{n_j}\right]^2. \tag{8.20}$$

### 8.5.2 Conditional Likelihood

The conditional distribution for $N = n_1$ is

$$\begin{aligned} f_{n_1}(n_1, k_{n_1}) &= f_{n_1}(k_{n_1})F(k_{n_1}), \\ f_{n_1}(n_1) &= \int f_{n_1}(k_{n_1})F(k_{n_1})dk_{n_1} = A_{n_1}(\mu), \\ f(k_{n_1}|n_1) &= \frac{f_{n_1}(k_{n_1})F(k_{n_1})}{A_{n_1}(\mu)}, \end{aligned}$$

from which the log-likelihood, score, Hessian, and information follow as:

$$\ell_{n_1}(\mu) = \ln\left[h_{n_1}(k_{n_1})\right] + \mu k_{n_1} - n_1\mu - \ln\left[A_{n_1}(\mu)\right], \tag{8.21}$$

$$S_{n_1}(\mu) = k_{n_1} - \frac{B_{n_1}(\mu)}{A_{n_1}(\mu)} = k_{n_1} - E\left[K\middle|N = n_1\right], \tag{8.22}$$

$$\begin{aligned} H_{n_1}(\mu) &= -\left\{\frac{C_{n_1}(\mu)}{A_{n_1}(\mu)} - \left(\frac{B_{n_1}(\mu)}{A_{n_1}(\mu)}\right)^2\right\} \\ &= -\left\{E\left[K^2\middle|N = n_1\right] - \left(E\left[K\middle|N = n_1\right]\right)^2\right\}, \\ I_{n_1}(\mu) &= E\left[K^2\middle|N = n_1\right] - \left\{E\left[K\middle|N = n_1\right]\right\}^2. \end{aligned} \tag{8.23}$$

Similarly for $N = n_j$ where $j > 1$, we have the conditional distribution:

$$f_{n_j}(n_j, k_{n_j}) = \widetilde{H}(k_{n_j})F(k_{n_j})\exp\left[\mu k_{n_j} - n_j a(\mu)\right], \tag{8.24}$$

$$f_{n_j}(n_j) = \int \widetilde{H}(k_{n_j})F(k_{n_j})\exp\left[\mu k_{n_j} - n_j a(\mu)\right] = A_{n_j}(\mu), \tag{8.25}$$

$$f(k_{n_j}|n_j) = \frac{\widetilde{H}(k_{n_j})F(k_{n_j})\exp\left[\mu k_{n_j} - n_j a(\mu)\right]}{A_{n_j}(\mu)}. \tag{8.26}$$

The following expressions for the likelihood, score, Hessian, and information are:

$$\ell_{n_j}(\mu) = \ln[\widetilde{H}(k_{n_j})F(k_{n_j})] + \mu k_{n_j} - n_j\mu - \ln A_{n_j}(\mu), \tag{8.27}$$

$$S_{n_j}(\mu) = k_{n_j} - \frac{B_{n_j}(\mu)}{A_{n_j}(\mu)} = k_{n_j} - E\left[K\middle|N = n_j\right], \tag{8.28}$$

$$\begin{aligned} H_{n_j}(\mu) &= -\left\{\frac{C_{n_j}(\mu)}{A_{n_j}(\mu)} - \left[\frac{B_{n_j}(\mu)}{A_{n_j}(\mu)}\right]^2\right\} \\ &= -\left\{E\left[K^2\middle|N = n_j\right] - \left(E\left[K\middle|N = n_j\right]\right)^2\right\}, \\ I_{n_j}(\mu) &= E\left[K^2\middle|N = n_j\right] - \left\{E\left[K\middle|N = n_j\right]\right\}^2. \end{aligned} \tag{8.29, 8.30}$$

The overall information for the conditional likelihood estimator is given by

$$\begin{aligned} I_c(\mu) &= \sum_{j=1}^{L} A_{n_j}(\mu)\left\{\frac{C_{n_j}(\mu)}{A_{n_j}(\mu)} - \left[\frac{B_{n_j}(\mu)}{A_{n_j}(\mu)}\right]^2\right\}, \\ &= \sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu) - \sum_{j=1}^{L} \frac{[B_{n_j}(\mu) - n_j a'(\mu) A_{n_j}(\mu)]^2}{A_{n_j}(\mu)}. \end{aligned} \tag{8.31}$$

From the scores (8.22) and (8.28), it can be seen that conditional likelihood estimator is unbiased. Clearly, the bias correction in the CLE mirrors the bias expression of the SA, as can be seen from (8.13). Upon writing (8.22) and (8.28), as

$$S_{n_j}(\mu) = k_{n_j} - n_j\mu + \left[n_j\mu - \frac{B_{n_j}(\mu)}{A_{n_j}(\mu)}\right],$$

the bias-correction factor in the CLE becomes even more apparent.

In contrast to the case of a fixed sample size, conditioning on the sample size in this case leads to loss of information, as can be seen by the subtraction of a positive factor in (8.31). This is a consequence of conditioning on a non-ancillary statistic, as discussed in Casella and Berger (2001).

For the CLE the MSE is:

$$\begin{aligned} \text{MSE}(\widehat{\mu}_c) &= \frac{1}{I_c(\mu)} \qquad (8.32) \\ &= \frac{1}{\sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu)} \\ &+ \frac{y}{\left[\sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu)\right]^2 - y\sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu)}, \qquad (8.33) \end{aligned}$$

where

$$y = \sum_{j=1}^{L} \frac{\left[B_{n_j}(\mu) - A_{n_j} n_j(\mu)\right]^2}{A_{n_j}(\mu)}.$$

The condition that $\text{MSE}(\widehat{\mu}) \geq \text{MSE}(\widehat{\mu}_c)$ is equivalent to the requirement that

$$\left[\sum_{j=1}^{L} \frac{[B_{n_j}(\mu) - A_{n_j} n_j \mu]}{n_j}\right]^2 \geq \frac{1}{\sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu)}$$

holds. For the special case of equal sample sizes this can never be true, hence the SA has the smaller MSE. More generally, neither is uniformly superior in terms of MSE.

## 8.6 Asymptotic Properties

We now turn to the large-sample properties of the estimators discussed in the previous sections. When $N \to \infty$, approximately $K \sim N(N\mu, N\sigma^2)$, so normal-theory arguments can be used. Considering a first-order Taylor series expansion of $F(k_{n_j})$ around $n_j\mu$ results in $F(k_{n_j}) \approx F(n_j\mu) + F'(n_j\mu)(k_{n_j} - n_j\mu)$. Without loss of generality, consider a class of stopping rules for which $F'(n_j) \overset{n\to\infty}{\to} 0$. In this setting, the expressions derived above can be approximated by

$$\begin{aligned}
A_{n_1}(\mu) &\approx F(n_1\mu), \\
B_{n_1}(\mu) &\approx F(n_1\mu)n_1\mu, \\
A_{n_j}(\mu) &\approx \prod_{i=1}^{j-1}[1 - F(n_i\mu)]F(n_j\mu), \quad (j > 1) \\
Bn_j(\mu) &\approx \prod_{i=1}^{j-1}[1 - F(n_i\mu)]F(n_j\mu)n_j\mu, \quad (j > 1).
\end{aligned}$$

These approximations will be useful in what follows.

### 8.6.1 Asymptotic Bias

Recall that the bias for the SA was given by (8.14), which asymptotically tends to the limit

$$\lim_{n\to\infty} \sum_{j=1}^{L} \frac{\prod_{i=1}^{j-1}[1 - F(n_i\mu)]F(n_j\mu)n_j\mu - \prod_{i=1}^{j-1}[1 - F(n_i\mu)]F(n_j\mu)n_j\mu}{n_j} \longrightarrow 0.$$

Although the sample average is finite-sample biased in general for data-dependent stopping rules, it is asymptotically unbiased and hence can be considered an appropriate candidate for practical use following a sequential trial. Emerson (1988) established the same result for two possible looks and further noted that this property is not relevant in group sequential trials, because large sample sizes are unethical, hence making the study of small sample properties crucial. On the other hand, results from a comprehensive analysis, comparing randomized controlled trials (RCTs) stopped for early benefit (truncated) and RCTs not stopped for early benefit (non-truncated), indicated that treatment effect was over-estimated in most of truncated RCTs regardless of the pre-specified stopping rule used (Bassler *et al.* , 2010). They further advocate stopping rules that demand large number of events. In the exploration of properties of estimators, in Chapter 7 it was shown that in the general class of linear

mean estimators, only the sample average has asymptotic unbiasedness property thus giving it an advantage in cases where asymptotic unbiasedness would play a role.

It has been noted previously that the bias reduces to zero when the conditional and marginal means are equal. We turn now to the asymptotic *conditional* behavior of the bias of the sample average *given the sample size*. Two cases are considered:

**Case I.** $F(n\mu) \overset{n\to\infty}{\longrightarrow} a \in ]0,1[$ and $F'(n\mu) \overset{n\to\infty}{\longrightarrow} 0$. For this case $E[\bar{\mu}|N = n_j] \overset{n\to\infty}{\longrightarrow} \mu$, for $j = 1, \ldots, L$.

**Case II.** Here, both the function $F(\cdot)$ and its first derivative $F(\cdot)$ converge to zero. When this happens, it does so for all but one of the sample sizes that can possibly be realized. The one exception is the sample size that will be realized, asymptotically, with probability one. Without loss of generality, we illustrate this case for stopping at the first look, assuming that the sample size realized at the first look corresponds to a set of values for $\mu$ taht do not contain the true one. Thus, $F(n\mu) \overset{n\to\infty}{\longrightarrow} 0$ and $F'(n\mu) \overset{n\to\infty}{\to} 0$. This case can correspond for particular forms of $F(k_{n_j})$. Given that $K$ is asymptotically normally distributed, letting $F(K) = \Phi(k)$ is a mathematically convenient choice from which it follows that $F(n_j\mu) = \Phi(n_j\mu)$. Consider first $N = n_1$. Then,

$$\lim_{n_1\to\infty} E[\bar{\mu}|N = n_1] = \mu - \lim_{n_1\to\infty} \frac{\phi(n_1\mu)\sigma^2}{\Phi(n_1\mu)},$$

of which the right hand term approaches $0/0$. We therefore apply l'Hopital's rule and obtain:

$$\lim_{n_1\to\infty} E[\bar{\mu}|N = n_1] = \mu - \lim_{n_1\to\infty} \frac{-n_1\mu\phi(n_1\mu)}{\phi(n_1\mu)} \to \infty,$$

with the sign opposite to that of $\mu$. Hence, conditional on the fact that stopping occurs after the first look, the estimate may grow in an unbounded way. However, recalling that $F(n\mu)$, the probability of stopping when $N = n_1$, also approaches zero, these extreme estimates are also extremely rare. In the same case, for $N = n_j$ $(j > 1)$, $\lim_{n\to\infty} E[\bar{\mu}|N = n_j] \to \mu$. So for these sample sizes no asymptotic bias occurs.

Recall from Chapter 7, that a large class of stopping rules corresponds to either Case I or Case II. For example, for stopping rule $\Phi(\alpha + \beta k/n)$, they found that Case I applies. Switching to $\Phi(\alpha + \beta k)$, $F'(n\mu) = \beta\phi(\alpha + \beta n\mu)$ which again tends to zero. However, $\Phi(\alpha + \beta n\mu)$ may tend to either zero or one. For a general rule $F(k) = \Phi(\alpha + \beta k n^m)$, with $m$ any real number, $F'(n\mu)$ converges to zero whatever $m$

is. Further, $F(n\mu)$ converges to $\Phi(\alpha+\beta\mu)$ for $m=-1$, $\Phi(\alpha)$ for $m<-1$, and $\Phi(\pm\infty)$ (i.e., 0 or 1) for $m>-1$. They showed that the sample average is asymptotically unbiased in all cases, and even conditionally asymptotically unbiased, based on the same logic as before. The above shows that this carries over to the case of an arbitrary number of looks.

### 8.6.2 Asymptotic Mean Square Error

Given that the bias for the sample average tends to zero as the sample size increases and that
$\sum_{j=1}^{L} B_{n_j}(\mu) - A_{n_j}(\mu) n_j \mu \overset{n\to\infty}{\longrightarrow} 0$, it follows that

$$\lim_{n\to\infty} \mathrm{MSE}(\widehat{\mu}) = \lim_{n\to\infty} \mathrm{MSE}(\widehat{\mu}_c) \to \frac{1}{\sum_{j=1}^{L} n_j a''(\mu) A_{n_j}(\mu)}.$$

## 8.7 Simulation Study

### 8.7.1 Design

The simulation study has been designed to corroborate the theoretical findings on the behavior of joint likelihood estimators, in comparison to commonly used biased adjusted estimators. Assume a clinical trial comparing a new therapy to a control, designed to follow O'Brien and Fleming's group sequential plan with four interim analyses.

The objective of the trial is to show that the mean response from the new therapy is higher than that of the control group. Let $Y_{it} \sim N(\mu_t, 1)$ and $Y_{ic} \sim N(\mu_c, 1)$ be the responses from subject $i$ in the therapy and control groups, respectively. The null hypothesis is formulated as $H_0 : \mu = \mu_t - \mu_c = 0$ vs. $H_1 : \mu = \mu_1 > 0$. Further, allow a type I error of 2.5% and 90% power to detect the clinically meaningful difference.

Given that we are interested in asymptotic behavior, different values of the clinically meaningful difference, $\mu_1 = 0.5, 0.25$, and $0.15$ are considered so as to achieve different sample sizes, with smaller $\mu_1$ corresponding to larger sample size.

With the settings described above, datasets are generated as follows; at each stage, $Y_{it} \sim N(2,1)$, $i = 1 \ldots n_j$, $j = 1 \ldots m$ $m = 4$ and $Y_{ic} \sim N(\mu_c, 1)$, where $\mu_c = 1.5$, 1.75, and 1.85 for the first, second, and third setting, respectively.

Estimation proceeds by obtaining the maximum likelihood estimator (sample average: $\widehat{\mu}_t - \widehat{\mu}_c$) at each stage and apply the stopping rule:

$$F(k_{n_j}) = \Phi\left(\alpha_j + \beta \frac{k_j}{n_j}\right), \qquad (j = 1 \ldots 4),$$

where $\beta = 100$ to represent the rules applied to the group sequential trials case as noted in Chapter 7. To follow O'Brien and Fleming boundaries, a value of $\alpha$ is chosen to make sure that the probability of stopping increases with the increase in number of looks, i.e., $\alpha_j = \frac{2(q-j+1)}{q}\alpha_1$, where $\alpha_1 = -50$, $-25$, and $-15$ for $\mu_1 = 0.5$, 0.25, and 0.15, respectively. $q$ is the number of planned looks. Obviously, the choice of $\alpha_j$ depends on the design and goals of the trial. In this setting, $\alpha_1$ was chosen such that $P(N = n_3|\mu = \mu_1) \geq 0.5$. The decision to stop is made when $F(k_{n_j}) > U$, where $U \sim \text{Uniform}(0, 1)$; otherwise, we continue.

The objective of the simulation is to show that the performance of the joint maximum likelihood estimator (MLE) as the mean estimator after a group sequential trial is as good as the other bias adjusted estimators, and that the confidence intervals obtained by using the observed information matrix, lead to valid conclusions.

The MLE will be compared against the median unbiased estimator (MUE), the bias adjusted estimator (BAM; Todd, Whitehead, and Facey 1996), and Rao's bias-adjusted estimator (RBADJ; Emerson and Fleming 1990).

Additional simulations with two possible looks and a smaller value of $\beta$ for both joint and conditional likelihood are presented in Appendix C.

### 8.7.2 Results

Table 8.1 gives the mean estimates for different estimators of $\mu$. On average MLE exhibits large relative bias compared to the bias adjusted estimates, for example, for $\mu_1 = 0.15$ which corresponds to maximum sample size of 1949, relative bias for MLE is 6% compared to 0.7% for BAM. On the other hand, MLE shows the asymptotic unbiasedness behaviour, seen by the reduction (though small) in relative bias as sample size increases. This is not the same for BAM and RBADJ.

While point estimates are useful in giving the picture of the magnitude of the treatment difference, confidence intervals (CI) are highly important in decision making. A comparison of adjusted confidence intervals provided with the RCTdesign package in R (Emerson *et al.*, 2012), to the likelihood based confidence intervals, obtained by using observed variance as precision estimates, indicates that their coverage probabilities are comparable. The coverage probabilities were (94.6%, 94.6%, 97.6%) for the adjusted CI and (93.8%, 92.8%, 96.8%) for MLE based CI, for the three settings in the order of increasing sample size. Using the same design parameters, we also investigated the type I error rate for MLE and adjusted estimators, by setting $\mu_1 = 0$ and obtaining the percentage times the confidence interval does not contain zero. Type I error rates for likelihood based CI were (5.6%,6.4%,2.8%), which are

similar to those based on adjusted CIs, (5.4%,4.8%,2.8%) for the three settings in the order of increasing sample size. Certainly using either of the CIs will lead to similar conclusions, which calls into question the necessity of adjusted methods in the analysis after group sequential trials, in line with the results derived.

We also explore the bias of each of the estimators at the sample level in contrast to the averaged bias as presented in Table 8.1. Recall that we had 500 samples for each setting, Table 8.2 gives the proportion of samples whose estimates' relative bias fell into a specified category. Indeed it is hard to pick a a preferred estimator based on these results since each of the estimator has about 75% of the estimates having relative bias of $> 10\%$. It is also clear from Figure 8.1, which plot the difference in relative bias, between each of the bias adjusted estimates and MLE, that none of the estimates discussed above is uniformly unbiased in comparison to MLE, i.e is some instances MLE may do better.

Table 8.1: *Mean estimates (Est.) and relative bias (R.Bias) for the three different settings of O'Brien and Fleming's design. Parameters common to all the three settings include, power=90%, type I error=0.025,* $H_0 : \mu = 0$ *vs.* $H_1 : \mu = \mu_1 > 0$*, where only the detectable difference (*$\mu_1$*) was changed to initiate change in maximum sample size (Size). MLE is the maximum likelihood estimate, BAM is the bias-adjusted maximum likelihood estimate, RBADJ is the Rao bias adjusted estimate and MUE is the median unbiased estimate.*

| | MLE | | BAM | | RBADJ | | MUE | |
|---|---|---|---|---|---|---|---|---|
| Size | Est. | R.Bias | Est. | R.Bias | Est. | R.Bias | Est. | R.Bias |
| 176 | 0.5448 | (0.0895) | 0.5142 | (0.0285) | 0.5019 | (0.0037) | 0.5251 | (0.0502) |
| 702 | 0.2665 | (0.0661) | 0.2508 | (0.0031) | 0.2473 | (0.0108) | 0.2557 | (0.0228) |
| 1949 | 0.1595 | (0.0635) | 0.1489 | (0.0070) | 0.1469 | (0.0209) | 0.1520 | (0.0130) |

## 8.8 Discussion

As a result of the bias associated with joint maximum likelihood estimators following sequential trials, much work has been applied to providing alternative estimators. The origin of the problem lies with the incompleteness of the sufficient statistic for the mean parameter Lehman (1950), implying, among others, that there is no best unbiased linear mean estimator.

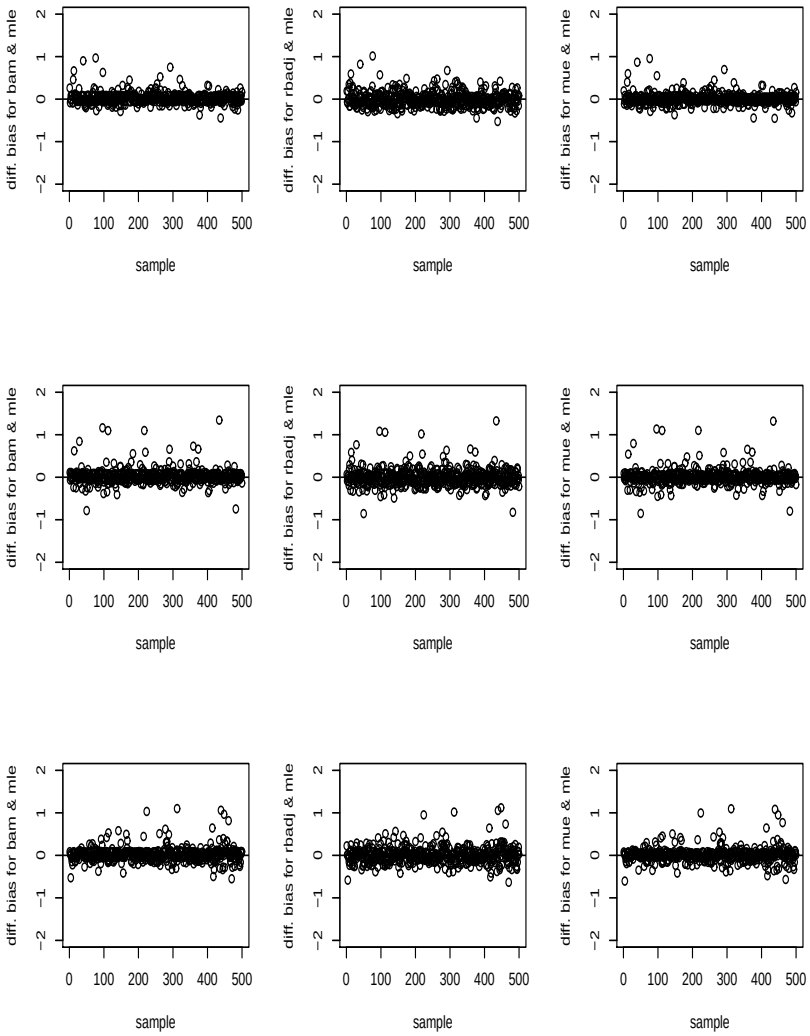


Figure 8.1: *Difference in relative bias between MLE and each the biased adjusted estimates (BAM, RBADJ and MUE). The first row is for* $\mu_1 = 0.5$*, second row,* $\mu_1 = 0.25$ *and third row,* |

Table 8.2: *Results from three different settings of O'Brien and Fleming's design. Parameters common to all three settings include: power=90%, type I error=0.025,* $H_0 : \mu = 0$ *vs.* $H_1 : \mu = \mu_1 > 0$*, where only the detectable difference* ($\mu_1$) *was changed to initiate change in maximum sample size (Size). Out of 500 datasets generated for each setting, we compare the proportion of estimates (Prop.) whose relative bias falls in the specified range (R.Bias). MLE is the maximum likelihood estimate, BAM is the biased adjusted maximum likelihood estimate, RBADJ is Rao's bias adjusted estimate, and MUE is the median unbiased estimate.*

| | | Prop. | | | |
|---|---|---|---|---|---|
| $\mu_1$(Size) | R.Bias | BAM | RBADJ | MUE | MLE |
| 0.5(176) | $\leq 0.99$ | 2.6 | 2.0 | 2.2 | 2.2 |
| | $1 - 4.99$ | 8.4 | 11.4 | 11.0 | 10.6 |
| | $5 - 10$ | 10.6 | 11.6 | 12.6 | 15.0 |
| | $> 10$ | 78.4 | 75.0 | 74.2 | 72.2 |
| 0.25(702) | $\leq 0.99$ | 2.0 | 3.2 | 1.4 | 2.6 |
| | $1 - 4.99$ | 7.2 | 9.0 | 8.8 | 9.0 |
| | $5 - 10$ | 9.4 | 9.8 | 10.8 | 9.8 |
| | $> 10$ | 81.4 | 78.0 | 79.0 | 78.6 |
| 0.15(1949) | $\leq 0.99$ | 2.6 | 1.8 | 1.4 | 2.2 |
| | $1 - 4.99$ | 7.4 | 13.2 | 8.0 | 13.2 |
| | $5 - 10$ | 9.2 | 9.0 | 11.8 | 11.0 |
| | $> 10$ | 80.8 | 76.0 | 77.6 | 74.4 |

Using stochastic stopping rules, which encompass the deterministic stopping rules used in sequential trials as special cases, we have studied the properties of joint maximum likelihood estimators afresh, in an attempt to enhance our understanding of the behaviour of estimators (for both bias and precision) based on data from such studies.

First, the incompleteness of the sufficient statistic when using a stochastic stopping rule has been established. Using a generalized sample average, it is noted that in almost no case is there an unbiased estimator. Even when such an estimator does exist, with a completely random sample size, it cannot be uniformly best.

Second, although for a sequential trial with a deterministic stopping rule, the ordinary sample average is finite sample biased, it can be been shown both directly and through likelihood arguments, that it is asymptotically unbiased and so remains a

good candidate for practical use. Further, it is computationally trivial, has a correspondingly simple estimator of precision, derived from observed information matrix and hence a well behaved asymptotic likelihood-based confidence interval. In addition, the mean square error of the sample average is smaller than that of the estimator based on the conditional likelihood, even though the latter is finite sample unbiased. The conditional estimator is also computationally more involved, because there is no closed-form solution. Asymptotically, the mean square errors of both estimators converge.

Third, there is the subtle issue that the sample average may be asymptotically biased for certain stopping rules, when its expectation is considered conditionally on certain values of the sample size. However, this is not a real practical problem because this occurs only for sample sizes that have asymptotic probability zero of being realized. We placed emphasis on joint and conditional likelihood estimators. While in the former the stopping rule is less present than sometimes thought, it is not in the latter. Also, when alternative frequentist estimators are considered, the stopping rule is likely to play a role in synchrony with the rule's influence on hypothesis testing due to the duality between hypothesis testing and confidence intervals.

While in some circumstances other sources of inaccuracy may overwhelm the issue studied here, we believe it is useful to bring forward implications of our findings for likelihood-based estimation.

In conclusion, the sample average is a very sensible choice for point, precision, and interval estimation following a sequential trial.

# Chapter 9

# Reliability Measures In Item Response Theory: Manifest *Versus* Latent Correlation Functions

Reliability of a scale of measurement is of paramount importance in measurement studies. For continuous scales the exact reliability measures, are easily obtained in contrast to categorical scales, where computational obstacles are rampant and approximations preferred instead. The commonly used approximation functions have drawbacks that limit their usability in some situations. In this chapter, we introduce new reliability approximation functions that overcome some of the limitations, elude the computational obstacles and provide close approximations to the true reliability. Though the implementation is to behavior and education psychology data, reliability of a scale, is also important in studies evaluating a drug for its impact on quality of life, and thus the results discussed in this chapter are relevant for clinical development of a drug.

## 9.1 Introduction

Measurement studies play a vital role in exploring various attributes, like social and intellectual behavior. The relevance of such studies depend on several, equally im-

portant factors, which include reliability of the tool used for measurement. Culligan (2008) defines reliability as a measure of the consistency of the application of an instrument to a particular population at a particular time. Classical test theory (CTT) remains one of the most used paradigms for the analysis of measurement studies, within which the concept of reliability is well developed. In CTT. the reliability measure is simply the proportion of true to observed variance. Limitations in classical theory (Schaeffer *et al.*, 1986; O'Brien, 1995) have led to development of several alternatives, in some of which deriving a reliability measure is straightforward, such as in CTT, while in others it is not.

Examples include generalizability theory (GT), which makes use of linear mixed models to estimate various variance components used in estimating reliability (Van Leeuwen et al., 1998). With linear mixed models, the definition of reliability as a proportion of true to observed variance is easily carried forward due to the nice properties of the normal distribution, which is usually assumed for the observed responses. The most notable being the separation of mean and variance parameters.

For binary response items, Item response theory (IRT) has indisputably commanded wide application among measurement studies, mostly for its advantages over classical theory (Rasch, 1960; De Ayala, 2009). Much as IRT models are commonly used for measurement of variables like attributes and attitudes (Van der Linden and Hambleton, 1997), the question of reliability of measurement (Spearman, 1904), which is crucial for such studies, cannot be ignored.

Unfortunately, peculiarities emerge when dealing with binary responses. For normally distributed outcomes, reliability of measurement reduces to, $\sigma_\theta^2/(\sigma_\theta^2+\sigma_\varepsilon^2)$, where $\sigma_\theta^2$ is the variance of the person trait $\theta$ and $\sigma_\varepsilon^2$ is the variance of the distribution assumed for the errors. This is commonly referred to as intraclass correlation (Molenberghs and Verbeke, 2005). Directly using intraclass correlation for dichotomous responses, produces what is known as *latent correlation* because it gives the correlation between responses at a logit or probit scale. It follows that reliability measured using latent correlation will be at a latent scale. More often than not, scientific interest is in the reliability of the observed scores rather than the latent ones such that meaningful reliability measures have to be based on *manifest correlation*, i.e., correlation between observed scores. While for normally distributed outcomes latent and manifest correlations coincide, this is no longer the case for binary, or other non-normal outcomes. Hence, to obtain the meaningful reliability measures, appropriate quantities for the intraclass correlation formula have to be derived.

Reliability measures based on manifest correlations are not as well developed for dichotomous responses as is the case for continous responses. Briggs and Wilson

(2007) and Rodríguez and Elo (2003) note that these are usually difficult to derive because they involve the evaluation of integrals that lack closed forms and thus are not widely used. To bypass such difficulties, approximate reliability measures are preferred and these include: Cronbach's $\alpha$, the intra-class correlation, and Fisher's Information measure. Drawbacks for using Cronbach's alpha in IRT, have been well documented (Cronbach and Shavelson, 2004), and Fisher's Information measure has limited application, given that under some conditions it can be negative (Mesbah *et al.*, 2002). Therefore it is not meaningful in some cases. Furthermore, its extension to models with multi-dimension traits is not clear. Some drawbacks regarding the use of the intraclass correlation in IRT, i.e., using reliability measures based on latent correlation, will be highlighted in the sections to follow.

Most of the IRT models fall into the family of generalized linear mixed model (GLMM), an extension of linear mixed models to a special family of non-linear mixed models (Molenberghs and Verbeke, 2005; Rijmen *et al.*, 2003), where the outcome is of a non-Gaussian type, but the effects of predictor variables still enter a so-called linear predictor function. Vangeneugden *et al.* (2010) derived approximate manifest variance-covariance functions and correlation functions for the GLMM family. The approximation totally evades the need to evaluate integrals and requires the input of estimates that are easily obtained during models estimation.

Taking into account that One- and Two-parameter Logistic (1PL and 2PL) models belong to the GLMM family, we explore the usefulness of such approximations in these two IRT models, towards obtaining reliability measures. Whereas the goal of Vangeneugden *et al.* (2010) was to estimate manifest correlations between two binary outcomes within a subject (which are the correlations of interest for most model members of the GLMM), by studying these approximation in the context of a combination of Classical Test Theory (CTT) and IRT, we derive reliability measures, both at the expected item score and expected sum score levels, that directly correspond to the definition of reliability in CTT, i.e., proportion of true to observed variance, which are based on manifest correlations. The performance of these measures will be assesed through a simulation study which will compare the newly derived approximations, latent and Fisher Information based reliability measures to the exact reliability. Applicability will be shown through an empirical data analysis of Verbal Aggression and Law School Admission Test datasets.

## 9.2 Reliability Measures in One Parameter Logistic (1PL) and Two Parameter Logistic (2PL) Models

Reliability measures of common interest for 1PL and 2PL models, which are also the focus of this work, include, reliability of expected item score and reliability of the expected sum score. This section reviews both exact and approximate methods for estimating such measures.

### 9.2.1 Exact Reliability Measures

Customarily, expected item and sum score reliability measures are computed based on observed scores, which are binary in nature for 1PL and 2PL models. While exact methods for estimating such measures are widely known, they are rarely used in practice because they are computationally intensive; instead, approximations are preferred. The exact measures are reviewed to facilitate comparison with the approximated one.

Define $Y_{ij}$ as a realized score on item $j = 1, \ldots, N$ by person $i = 1, \ldots, I$. It is common to express the observed score as:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij},$$

where $\mu_{ij}$ is the true score and $\varepsilon_{ij}$ is the error score. Equivalently, the 2PL model formulates the observed score as

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij} = \frac{\exp\left[\alpha_j(\theta_i - \beta_j)\right]}{1 + \exp\left[\alpha_j(\theta_i - \beta_j)\right]} + \varepsilon_{ij}, \tag{9.1}$$

where $\theta_i$ is the person trait score, with $\theta_i$ having variance $\sigma_\theta^2$, $\alpha_j$ and $\beta_j$ are the discrimination and difficulty values for item $j$, and $\varepsilon_{ij}$ the error term, which in this case is a function of $\mu_{ij}$.

Recall that our focus is on the expected item and expected sum scores, which are given by:

$$\begin{aligned} Y_j &= \int \left\{ \frac{\exp\left[\alpha_j(\theta_i - \beta_j)\right]}{1 + \exp\left[\alpha_j(\theta_i - \beta_j)\right]} + \varepsilon_{ij} \right\} \phi(\theta_i|0, \sigma_\theta^2)\, d\theta \;\; = \mu_j + \varepsilon_j \\ S_T &= \int \sum_{j=1}^{N} \left\{ \frac{\exp\left[\alpha_j(\theta_i - \beta_j)\right]}{1 + \exp\left[\alpha_j(\theta_i - \beta_j)\right]} + \varepsilon_{ij} \right\} \phi(\theta_i|0, \sigma_\theta^2)\, d\theta \;\; = \mu + \varepsilon, \end{aligned} \tag{9.2}$$

respectively. Dimitrov (2003) defines their corresponding variances as:

$$\begin{aligned}
\text{var}(\varepsilon_j) &= \int \mu_{ij}(1-\mu_{ij})\,\phi(\theta_i|0,\sigma_\theta^2)\,d\theta, \\
\text{var}(\mu_j) &= \mu_j(1-\mu_j) - \text{var}(\varepsilon_j), \\
\text{var}(\varepsilon) &= \sum_{j=1}^{N} \text{var}(\varepsilon_j), \\
\text{var}(\mu) &= \int \left\{ \sum_{j=1}^{N} \frac{\exp\left[\alpha_j(\theta_i-\beta_j)\right]}{1+\exp\left[\alpha_j(\theta_i-\beta_j)\right]} \right\}^2 \phi(\theta_i|0,\sigma_\theta^2)\,d\theta \\
&\quad + \left\{ \int \sum_{j=1}^{N} \frac{\exp\left[\alpha_j(\theta_i-\beta_j)\right]}{1+\exp\left[\alpha_j(\theta_i-\beta_j)\right]} \phi(\theta_i|0,\sigma_\theta^2)\,d\theta \right\}^2,
\end{aligned}$$

with reliability defined as the proportion of true variance to observed variance. Further,

$$\begin{aligned}
\rho_i &= \frac{\text{var}(\mu_j)}{\text{var}(\mu_j)+\text{var}(\varepsilon_j)}, \\
\rho_s &= \frac{\text{var}(\mu)}{\text{var}(\mu)+\text{var}(\varepsilon)}
\end{aligned}$$

measure the expected item score and expected sum score reliability, respectively. Each of the components for obtaining these measures involves integration of normal random effects over binary data distributions, which are known to lack a closed form. The computational burden associated with these measures arguably results into their infrequent use in practice.

### 9.2.2 Intra-class Correlation (Latent)

The definition of the intra-class correlation stems directly from the definition of reliability in CTT, which is the ratio of the true over the observed variance. Let $Y_{ij}$ be a continuous observed score for person $i = 1, \ldots, I$ on item $j = 1, \ldots, N$, and further $\boldsymbol{Y}_i^c = (Y_{i1}, \ldots, Y_{ij})$. Using a linear mixed model, the observed score can be expressed in terms of the true score $\boldsymbol{\mu}_i$ and error as follows:

$$\boldsymbol{Y}_j^c = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\theta}_i - \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{9.3}$$

where $\boldsymbol{\beta}$ is the vector of item difficulties, $\theta_i \sim N(0, \sigma_\theta^2)$, and $\boldsymbol{\varepsilon}_i \sim N(0, \sigma_\varepsilon^2 \mathbf{I}_I)$, are the measures of person trait and random errors, respectively. Further, $\text{Cov}(\theta_i, \boldsymbol{\varepsilon}_i) = 0$. It has been shown that

$$\text{Var}(\boldsymbol{Y}_i^c) = \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) = \mathbf{1}_I \sigma_\theta^2 \mathbf{1}'_I + \sigma_\varepsilon^2 \mathbf{I}_I$$

(Verbeke and Molenberghs, 2000). For illustrative purposes, consider the case of $I = 2$. Then,

$$\text{Var}(\boldsymbol{\mu}_i) = \begin{bmatrix} \sigma_\theta^2 & \sigma_\theta^2 \\ \sigma_\theta^2 & \sigma_\theta^2 \end{bmatrix}, \qquad \text{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix}. \tag{9.4}$$

Item and sum score reliability measures follow as:

$$\rho_{i_1} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}, \quad \rho_{s_2} = \frac{2\sigma_\theta^2}{2\sigma_\theta^2 + \sigma_\varepsilon^2}, \tag{9.5}$$

which correspond to ICC(1) and ICC(k) measures in McGraw and Wong (1996).

The same arguments can be followed when $\boldsymbol{\mu}_i = \alpha(\boldsymbol{\theta}_i - \boldsymbol{\beta})$ and the equivalent quantities are:

$$\begin{aligned}
\text{Var}(\boldsymbol{Y}_i^c) &= \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\alpha}_I \sigma_\theta^2 \boldsymbol{\alpha}_I' + \sigma_\varepsilon^2 \mathbf{I}_I, \\
\text{Var}(\boldsymbol{\mu}_i) &= \begin{bmatrix} \alpha_j^2 \sigma_\theta^2 & \alpha_j \alpha_{i'} \sigma_\theta^2 \\ \alpha_j \alpha_{j'} \sigma_\theta^2 & \alpha_{i'}^2 \sigma_\theta^2 \end{bmatrix}, \\
\text{Var}(\boldsymbol{\varepsilon}_i) &= \begin{bmatrix} \sigma_\varepsilon^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix}, \\
\rho_{i_2} &= \frac{\alpha_j^2 \sigma_\theta^2}{\alpha_j^2 \sigma_\theta^2 + \sigma_\varepsilon^2}, \qquad (9.6) \\
\rho_{s_2} &= \frac{\sigma_\theta^2 (\alpha_j + \alpha_{j'})^2}{\sigma_\theta^2 (\alpha_j + \alpha_{j'})^2 + 2\sigma_\varepsilon^2}. \qquad (9.7)
\end{aligned}$$

Now, consider a binary observed score modeled through the 1PL model as:

$$\text{logit}\left[P(\boldsymbol{Y}_j = 1)\right] = \boldsymbol{\theta}_i - \boldsymbol{\beta}. \tag{9.8}$$

It has been noted (Agresti, 2002; Rodríguez and Elo, 2003) that (9.8) can be expressed in the form of (9.3), as follows:

$$\boldsymbol{Y}_j^* = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i = \boldsymbol{\theta}_i - \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^*, \tag{9.9}$$

where $Y_{ij}^*$ is assumed to be a latent continuous score underlying the dichotomization of $Y_{ij}$, such that $Y_{ij} = 1$ if $Y_{ij}^* \geq C$ and 0 otherwise, with $C$ a pre-specified threshold and $\varepsilon^* \sim \text{Logistic}(0, 1)$. Model (9.9) is a typical linear mixed model, hence the theory behind the derivation of reliability measures (9.5), (9.6), and (9.7), can be applied directly and the corresponding reliability measures are:

$$\rho_{i_{l1}} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \frac{\pi^2}{3}} \quad \text{and} \quad \rho_{s_{l1}} = \frac{2\sigma_\theta^2}{2\sigma_\theta^2 + \frac{\pi^2}{3}}, \tag{9.10}$$

for item and test score, respectively, for the 1PL model. Equivalent expressions for the 2PL model are:

$$\rho_{i_{l2}} = \frac{\alpha_j^2\sigma_\theta^2}{\alpha_j^2\sigma_\theta^2 + \frac{\pi^2}{3}}, \quad \text{and} \quad \rho_{s_{l2}} = \frac{\sigma_\theta^2(\alpha_j + \alpha_{j'})^2}{\sigma_\theta^2(\alpha_j + \alpha_{j'})^2 + \frac{2\pi^2}{3}}, \tag{9.11}$$

for the item and test reliability, respectively, $\pi^2/3$ is the variance of the underlying error distribution, the standard logistic density. While the mathematical motivation is appealing, it is evident that (9.10) and (9.11) only depend on $\sigma_\theta^2$. Hence,‘ any change in this value, for example, due to change in identification restrictions, will result in varying reliability. These are examples of reliability measures based on latent correlation and, as noted before, for the binary case such measures do not coincide with their manifest correlation based counterparts. In a way this is unfortunate, especially in cases where there are no closed forms for the marginal model stemming from the hierarchical formulation. Note that the difference between latent and manifest correlation is independent of the existence of a closed form. For example, a probit link with normal random effects allows for a closed-form formulation (Molenberghs *et al.* , 2010), but also there the two correlations have a different expression. Depending on the research question, either manifest or latent or both correlations can be of interest, in spite of the fact that manifest correlation is more difficult to work with than latent correlation, latent correlation should not be used when manifest correlation is of interest.

### 9.2.3 Fisher Information

$$I(\theta_i; I) = \sum_{i=1}^{I} \alpha_j^2 \frac{\exp(\eta_{ij})}{[1 + \exp(\eta_{ij})]^2},$$

with $\eta_{ij} = \alpha_j(\theta_i - \beta_j)$, where $\beta_j$ is the difficulty value for item $i$, $\theta_i$ is the value of the latent trait for person $j$, and $I$ is the total number of items. The expression is intuitively appealing, given that it is a sum over standard logistic variances. An approximation of the reliability coefficient follows as:

$$\rho_f = 1 - \frac{1}{N}\sum_{j=1}^{N} \frac{I(\theta_i; I)^{-1}}{\sigma_{\hat{\theta}}^2} \tag{9.12}$$

(Lord, 1983), where $\sigma_{\hat{\theta}}^2$ is the variance of the estimated latent trait, e.g., the observed variance of person parameters and $N$ the total number of persons. This approximation is only valid when the number of items is large; it also requires the knowledge of true values of difficulty parameters, information which usually is beyond reach. A 1PL equivalent follows by setting $\alpha = 1$.

## 9.3 Taylor-series-based Derivation of the Correlation Function

### 9.3.1 Manifest Correlation Functions For GLMM

We present here a brief review of the explicit but approximate correlation functions, as based on Taylor-series approximations as derived for GLMM family and explained in detail in Vangeneugden *et al.* (2010). Let $Y_{ij}$ be the $j^{th}$ outcome measured on person $i$ , $j = 1, \ldots N$ and $i = 1, \ldots, I$; further let $\boldsymbol{Y}_i = Y_{i1}, \ldots, Y_{iN}$.

Write the general model as $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, where the conditional mean, given the random effects are written as $\boldsymbol{\mu}_j = h(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\theta}_i)$, and where $\boldsymbol{\beta}$ is a vector of fixed effects parameters, $\boldsymbol{\theta}_i$ are random effects, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are known design matrices, and $h$ is a known link function. Finally, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{iN})'$ is the residual error component.

The general formula for the variance-covariance matrix of $\boldsymbol{Y}_i$ was derived without any restriction on the distribution of the outcome variable, nor on the complexity of the model, e.g., allowing for serial correlation or not. This maximizes the similarity with the case of continuous, normally distributed outcomes. However, a key distinction is that in the Gaussian case the mean and variance are functionally independent parameters, whereas here the residual variance will follow from the mean. The variance-covariance matrix can be written as:

$$\boldsymbol{V}_i = \text{Var}(\boldsymbol{Y}_i) = \text{Var}(\boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i) = \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i) + 2\text{Cov}(\boldsymbol{\mu}_i, \boldsymbol{\varepsilon}_i). \tag{9.13}$$

Because $\mu_i$ depends on $\boldsymbol{\theta}_i$ only, which is independent of $\boldsymbol{\varepsilon}_i$, it follows that $\text{Cov}(\mu_i, \boldsymbol{\varepsilon}_i) = 0$, and the first term in (9.13), using a first-order Taylor series expansion around $\boldsymbol{\theta}_i = 0$ reduces to:

$$\begin{aligned} \text{Var}(\boldsymbol{\mu}_i) &= \text{Var}[\boldsymbol{\mu}_i(\boldsymbol{\eta}_i)] = \text{Var}[\boldsymbol{\mu}_i(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\theta}_i)] \\ &\cong \left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\eta}_i}\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\theta}_i}\mid_{\boldsymbol{\theta}_i=0}\right)\boldsymbol{D}\left(\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\eta}_i}\frac{\partial\boldsymbol{\eta}_i}{\partial\boldsymbol{\theta}_i}\mid_{\boldsymbol{\theta}_i=0}\right)' \cong \boldsymbol{\Delta}_i\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i'\boldsymbol{\Delta}_i', \end{aligned} \tag{9.14}$$

where $\boldsymbol{\Delta}_i = \frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\eta}_i}\mid_{\boldsymbol{\theta}_i=0}$. The second term in (9.13), leads to:

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \text{Var}[E(\boldsymbol{\varepsilon}_i \mid \boldsymbol{\theta}_i)] + E[\text{Var}(\boldsymbol{\varepsilon}_i|\boldsymbol{\theta}_i)] = E[\text{Var}(\boldsymbol{\varepsilon}_i \mid \boldsymbol{\theta}_i)] = \boldsymbol{\Xi}_i^{\frac{1}{2}}\boldsymbol{\Sigma}_i\boldsymbol{\Xi}_i^{\frac{1}{2}}, \tag{9.15}$$

where $\boldsymbol{\Xi}$ is a diagonal matrix with the overdispersion parameters along the diagonal. If there are no overdispersion parameters, $\boldsymbol{\Xi}$ is set equal to the identity matrix. Expand the variance function $\boldsymbol{\Sigma}_i$ so that

$$\text{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Xi}_i^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{R}_i\boldsymbol{\Xi}_i^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}, \tag{9.16}$$

where $\boldsymbol{R}_i$ is the correlation matrix and $\boldsymbol{A}_i$ is a diagonal matrix containing the variances following from the generalized linear model specification of $Y_{ij}$ given the random effects $\boldsymbol{\theta}_i = 0$, i.e., with diagonal elements $v(\mu_{ij})|_{\boldsymbol{\theta}_i=0}$. Using (9.14) and (9.16), we have the following expression for the variance-covariance matrix (9.13):

$$\boldsymbol{V}_i = \boldsymbol{\Delta}_i \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' \boldsymbol{\Delta}_i' + \boldsymbol{\Xi}_i^{\frac{1}{2}} \boldsymbol{A}_i^{\frac{1}{2}} R_i \boldsymbol{\Xi}_i^{\frac{1}{2}} \boldsymbol{A}_i^{\frac{1}{2}}. \tag{9.17}$$

Evidently, from this variance-covariance matrix, we can easily obtain the correlations. While the above derivation is referred to as a first-order Taylor series expansion, the exact same expression follows if a second-order expansion is considered, owing to terms vanishing. Therefore, we are authorized to refer to it as a second-order Taylor series expansion which according to Vangeneugden *et al.* (2011), who explored the quality of approximation by considering higher-order series, gives a good approximation.

### 9.3.2 Taylor Series Based Reliability Measures For 1PL and 2PL Models.

Loosely, reliability is an indicator of strength of agreement of particular scores (depending on the form of reliability); usually it takes the form of a correlation function. To obtain reliability measures on the scale of observed scores, manifest correlation functions have to be used. While such functions are easily obtained for continuous data, usually for binary data, approximations are employed. This section digests the relevance of the approximate variance-covariance matrix in a GLMM for assessing reliability of the expected item score and the expected sum score.

Without loss of generality, consider a measurement tool with $j = 1, \ldots, N$ items, responded to by $i = 1, \ldots, I$ persons and further let $Y_{ij}$ be the binary score on item $j$ by person $i$. Then, parameterize the 2PL model as

$$Y_{ij} = \underbrace{\frac{\exp\left[\alpha_j(\theta_i - \beta_j)\right]}{1 + \exp\left[\alpha_j(\theta_i - \beta_j)\right]}}_{\mu_{ij}} + \varepsilon_{ij}, \tag{9.18}$$

where $\theta_i \sim N(0, \sigma_\theta^2)$, $\beta_j$ and $\alpha_j$ are the difficulty value and discrimination parameters, respectively, for item $i$. The model formulation in Section 9.3.1, $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, is basically the matrix representation of (9.18), where $\boldsymbol{Y}_i$ is the vector of $Y_{ij}$'s, for person $j$ on all items; we proceed similarly for $\boldsymbol{\mu}_i$ and $\boldsymbol{\varepsilon}_i$.

Vangeneugden *et al.* (2010) approximate $\text{Var}(\boldsymbol{Y}_i)$ by using a first-order Taylor series expansion of the variance function around $\theta_i = 0$. Implicitly, this assumes that $P(y_{ij} = 1|\theta_i) = P(y_i = 1)$, which defines the expected item score (Section 9.2.1).

Further, $\boldsymbol{Y}_i = \boldsymbol{Y}_{j'}$, $j \neq j'$ and $\boldsymbol{Y}_i = \boldsymbol{Y}_i$, where $\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_i$, is the vector of expected item scores. Consequently,

$$\begin{aligned} \text{var}(\boldsymbol{Y}_i) = \text{var}(\boldsymbol{Y}_i) &\cong \text{Var}(\boldsymbol{\mu}_i) + \text{Var}(\boldsymbol{\varepsilon}_i), \quad \text{where} && (9.19)\\ \text{Var}(\boldsymbol{\mu}_i) &\cong \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_i}\,|_{\boldsymbol{\theta}_i=0}\right)\boldsymbol{D}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\eta}_i}\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_i}\,|_{\boldsymbol{\theta}_i=0}\right)', && (9.20)\\ \text{Var}(\boldsymbol{\varepsilon}_i) &\cong \boldsymbol{A}_i^{\frac{1}{2}}\boldsymbol{A}_i^{\frac{1}{2}}. && (9.21) \end{aligned}$$

The fact that $\boldsymbol{R}_i$ and $\boldsymbol{\Xi}_i$ disappear is a result of conditional independence and the assumption of no overdispersion in the 2PL model. From (9.2), it is easy to deduce that the expected sum score merely is the sum over all expected item scores. Hence, the variance of $S_T$ is just the sum of all components in $\text{Var}(\boldsymbol{Y}_i)$, i.e.,

$$\text{Var}(S_T) = \sum \text{Var}(\boldsymbol{Y}_i) = \sum \text{Var}(\boldsymbol{\mu}_i) + \sum \text{Var}(\boldsymbol{\varepsilon}_i).$$

Using the classical definition of reliability, i.e., the proportion of true to observed variance, we obtain:

$$\rho_{iA} = \frac{\text{Var}(\mu_i)}{\text{Var}(u_i) + \text{Var}(\varepsilon_i)} \quad \text{and} \quad \rho_{sA} = \frac{\text{Var}(\sum_i \boldsymbol{\mu}_{ij})}{\text{Var}(\sum_i \boldsymbol{\mu}_{ij}) + \text{Var}(\sum_i \boldsymbol{\varepsilon}_{ij})}, \tag{9.22}$$

as reliability measures for the expected item and test scores, respectively. $\text{Var}(\mu_i)$ and $\text{Var}(\varepsilon_i)$ correspond to the $i^{th}$ diagonal elements of $\text{Var}(\boldsymbol{\mu}_i)$ and $\text{Var}(\boldsymbol{\varepsilon}_i)$, respectively. Equivalent expressions for the 1PL model follow when $\alpha_j = 1$.

It is assumed that variance estimation comes after estimation of other model parameters. As such, the assumptions made in this section apply to variance estimation only.

We acknowledge alternative approximations for reliability measures based on manifest correlations (Dimitrov, 2003). Nevertheless, the simplicity and easy-to-follow nature of our approximations make them a valuable addition to the existing methods.

#### 9.3.2.1 Illustration For 1PL Model

This section illustrates computation of (9.19) and (9.22) for 1PL, a step that also allows the exploration of differences in latent and manifest correlations.

Define the 1PL model as (9.18), where $\alpha_j = 1$. It follows that:

$$\begin{aligned} &\eta_{ij} = \theta_i - \beta_j, \qquad \frac{\partial \eta_{ij}}{\partial \theta_j}|_{\theta_i=0} = 1, && (9.23)\\ &\frac{\partial \mu_{ij}}{\partial \eta_{ij}}|_{\theta_i=0} = \frac{\exp(\beta_j)}{[1+\exp(\beta_j)]^2} = v(\mu_{ij})|_{\boldsymbol{\theta}_i=0}. \end{aligned}$$

Assume a test with only two items, then,

$$\frac{\partial \boldsymbol{\mu}_{ij}}{\partial \boldsymbol{\eta}_{ij}}|_{\theta_i=0} = \boldsymbol{A}_i = \begin{bmatrix} v_{ij}(0) & 0 \\ 0 & v_{ij'}(0) \end{bmatrix}, \quad \boldsymbol{D} = \left[\sigma_\theta^2\right], \quad \frac{\partial \boldsymbol{\eta}_{ij}}{\partial \theta_j}|_{\boldsymbol{\theta}_i=0} = 1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad (9.24)$$

$$\text{Var}(\boldsymbol{\mu}_i) = \begin{bmatrix} v_{ij}^2(0)\sigma_\theta^2 & v_{ij}(0)v_{ij'}(0)\sigma_\theta^2 \\ v_{ij'}(0)v_{ij}(0)\sigma_\theta^2 & v_{ij'}^2(0)\sigma_\theta^2 \end{bmatrix}, \quad \text{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} v_{ij}(0) & 0 \\ 0 & v_{ij'}(0) \end{bmatrix},$$

$$\rho_{i_{A1}} = \frac{v_{ij}(0)\sigma_\theta^2}{1 + v_{ij}(0)\sigma_\theta^2}, \qquad \rho_{s_{A1}} = \frac{\sigma_\theta^2\left[v_{ij}(0) + v_{ij'}(0)\right]}{1 + \sigma_\theta^2\left[v_{ij}(0) + v_{ij'}(0)\right]},$$

where $v_{ij}(0) = v(\mu_{ij})|_{\boldsymbol{\theta}_i=0}$.

**Manifest versus Latent Correlation**

The latent correlation measure presented in Section 9.2.3 is obviously appealing and easy to obtain, and one can be tempted to use it as a reliability measure. We study the relationship between the reliability measures based on latent and manifest correlation.

Consider $\rho_{i_{A1}}$ and $\rho_{s_{A1}}$ at their maximum possible values, which are easily obtained by realizing that the maximum value for $v_{ij}(0) = 0.25$. A comparison of these to the latent reliability measures in (9.10), reveals the following relationship;

$$\rho_{i_{A1}} = \frac{\sigma_\theta^2}{4 + \sigma_\theta^2} < \rho_{i_{l1}} = \frac{\sigma_\theta^2}{\frac{\pi^2}{3} + \sigma_\theta^2}, \quad (9.25)$$

$$\rho_{s_{A1}} = \frac{\sigma_\theta^2}{2 + \sigma_\theta^2} < \rho_{s_{l1}} = \frac{\sigma_\theta^2}{\frac{\pi^2}{6} + \sigma_\theta^2}. \quad (9.26)$$

Latent correlation based score reliability is always greater than its manifest correlation based counterpart, hence if scientific interest is on reliability of observed scores, great caution has to be exercised in using latent correlation based reliability measures.

#### 9.3.2.2 Illustration For 2PL Model

Similar to Section (9.3.2.1) we illustrate computation of (9.19) and (9.22) for the 2PL model, and further explore the relationship between latent and manifest correlations.

Define the 2PL model as (9.18), it follows that,

$$\eta_{ij} = \alpha_j(\theta_i - \beta_j), \qquad \frac{\partial \eta_{ij}}{\partial \theta_j}|_{\theta_i=0} = \alpha_j \quad (9.27)$$

$$\frac{\partial \mu_{ij}}{\partial \eta_{ij}}|_{\theta_i=0} = \frac{\exp(\alpha_j\beta_j)}{[1 + \exp(\alpha_j\beta_j)]^2} = v(\mu_{ij})|_{\boldsymbol{\theta}_i=0}.$$

Again consider a test with only two items, then,

$$\frac{\partial \boldsymbol{\mu}_{ij}}{\partial \boldsymbol{\eta}_{ij}}|_{\theta_i=0} = A_i = \begin{bmatrix} v_{ij}(0) & 0 \\ 0 & v_{j'i}(0) \end{bmatrix}, \quad D = \left[\sigma_\theta^2\right], \qquad \frac{\partial \boldsymbol{\eta}_{ij}}{\partial \theta_j}|_{\boldsymbol{\theta}_i=0} = \begin{bmatrix} \alpha_j \\ \alpha_{j'} \end{bmatrix}, \quad (9.28)$$

$$\mathrm{Var}(\boldsymbol{\mu}_i) = \begin{bmatrix} \alpha_i^2 v_{ij}^2(0)\sigma_\theta^2 & \alpha_{j'}\alpha_j v_{ij}(0)v_{ij'}(0)\sigma_\theta^2 \\ \alpha_{j'}\alpha_j v_{ij'}(0)v_{ij}(0)\sigma_\theta^2 & \alpha_{i'}^2 v_{ij'}^2(0)\sigma_\theta^2 \end{bmatrix},$$

$$\mathrm{Var}(\boldsymbol{\varepsilon}_i) = \begin{bmatrix} v_{ij}(0) & 0 \\ 0 & v_{ij'}(0) \end{bmatrix},$$

$$\rho_{i_{A2}} = \frac{v_{ij}(0)\alpha_i^2\sigma_\theta^2}{1 + v_{ij}(0)\alpha_i^2\sigma_\theta^2}, \quad \rho_{s_{A2}} = \frac{\sigma_\theta^2\left[v_{ij}(0)\alpha_j + v_{ij'}(0)\alpha_{j'}\right]^2}{\left[v_{ij}(0) + v_{ij'}(0)\right] + \sigma_\theta^2\left[v_{ij}(0)\alpha_j + v_{ij'}(0)\alpha_{j'}\right]^2},$$

where $v_{ij}(0) = v(\mu_{ij})|_{\boldsymbol{\theta}_i=0}$.

**Manifest versus Latent Correlation.**

The following relationship exists between latent and manifest correlation based reliability measures for the 2PL model:

$$\rho_{i_{A2}} = \frac{\alpha_j^2\sigma_\theta^2}{4 + \alpha_j^2\sigma_\theta^2} \quad < \quad \rho_{i_{l2}} = \frac{\alpha_j^2\sigma_\theta^2}{\frac{\pi^2}{3} + \alpha_j^2\sigma_\theta^2}, \quad (9.29)$$

$$\rho_{s_{A2}} = \frac{\sigma_\theta^2(\alpha_j + \alpha_{j'})^2}{8 + \sigma_\theta^2(\alpha_j + \alpha_{j'})^2} \quad < \quad \rho_{s_{l2}} = \frac{\sigma_\theta^2(\alpha_j + \alpha_{j'})^2}{\frac{2\pi^2}{3} + \sigma_\theta^2(\alpha_j + \alpha_{j'})^2} \quad (9.30)$$

The relationship between latent and manifest correlations based reliability measures observed in the 1PL model is the same for the 2PL model.

## 9.4 Simulation Study

Due to lack of closed form quantities, the performance of reliability measures based on Taylor series approximations of the variance-covariance matrix in Section 9.3.2, will be assessed through a simulation study. We will compare them with the exact measures described in Section 9.2.1. Additionally, the relationship observed between manifest and latent correlation based measures in Sections 9.3.2.1 and 9.3.2.2, will be studied for more than two items.

### 9.4.1 Design of the Simulation Study

Measuring tools calibrated under the 1PL and 2PL models are considered. Each has 24 items, whose difficulty values ($\beta_j$) are generated from a uniform distribution within the range $[-4, 4]$, the discrimination parameter values ($\alpha_j$) for the 2PL model are sampled from $N(2, 0.64)$, and the number of respondents is set to 400. To assess the quality of our approximations at different levels of reliability, three values for the variance of person trait scores ($\theta_i$), which influences reliability, were considered, i.e., $\sigma_\theta^2 = (0.25, 1, 4)$, implying that each model will produce three measuring tools that will be responded to by three different sets of individuals. With these values, the score $Y_{ij}$ for item $j$ by person $i$ is generated from a Bernoulli($\pi$), where

$$\pi = \begin{cases} \frac{\exp(\theta_i - \beta_j)}{1+\exp(\theta_i - \beta_j)} & \text{for the 1PL model,} \\ \frac{\exp[\alpha_j(\theta_i - \beta_j)]}{1+\exp[\alpha_j(\theta_i - \beta_j)]} & \text{for the 2PL model.} \end{cases}$$

For the exact, Taylor series approximation and latent correlation based approximations, both expected item reliability and expected sum score reliability will be computed, while for the Fisher information coefficient, only the expected sum score will be obtained. To concentrate on the performance of the reliability measures, the models are not fit. Rather, we assume that the generated samples represent the true population. As such, simulated values are plugged into all the formulas and the integrals in Section 9.2.1 are obtained by simply averaging over the relevant quantities. This helps to eliminate behavior that may be observed due other issues, like lack of model convergence resulting into poor estimates.

### 9.4.2 Simulation Results

Table 9.1 indicates that the reliability of the expected sum, estimated using the newly introduced, Taylor series approximation is very similar to the exact reliability. Practically, using one in place of the other should lead to virtually the same substantive conclusions. In addition, Taylor series approximations are not computationally intensive, because they do not involve the evaluation of integrals with no closed forms, and quantities for approximating the variance-covariance matrix follow directly from model estimation. Clearly, these strengths make the Taylor series approximation a valuable addition to the theory of reliability measurement.

Also revealed in Table 9.1 is a possible drawback for the commonly used Fisher information based measure: when $\sigma_\theta^2 = 0.25$, reliability is negative and does not have a meaningful interpretation. Further, in other cases, like for the 2PL model with $\sigma_\theta^2 = 1$, it overestimates the exact reliability while in the 1PL model with $\sigma_\theta^2 = 0.25$,

it underestimates the exact reliability. In general, Fisher information based reliability will not always yield a truthful picture of the reliability of the expected sum.

In line with theory, latent correlation based reliability is consistently greater than the manifest correlation based measures (Taylor approximation and exact). If reliability at the logit scale is of interest, latent correlations are meaningful; otherwise, they should be avoided.

Even though maximum information is the target when creating item banks, expected item score reliability can also be useful, as it can help in choosing items that are reliable. Results in Table 9.2 emphasize why latent correlation based reliability may not be best suited to be used for such a process, especially for the 1PL model, because not only is it always greater than the manifest correlation based version, but it is also constant for all items. The Taylor series based versions approximate the exact expected item score reliability closely for the 1PL model, such that decision making based on the former is likely to reflect the decisions that would result from using the latter. For the 2PL model, Taylor series approximation is not good for some items like 3, 4, 8, 9, 11, 13, 15, and 21, which greatly underestimate the exact reliability, suggesting the need to improve the approximations if they are to be used for item reliability in the 2PL model. Possible considerations include expansion of Taylor series around $\theta_i = \widehat{\theta}_i$, the maximum likelihood estimate of the trait score, instead of $\theta_i = 0$, but this is beyond the scope of the current work. Latent reliability in 2PL increases with discriminative power, i.e., items with high discriminative power have high reliability, regardless of the difficulty level. Results for cases with $\sigma_\theta^2 = 0.25, 4$ are presented in the Appendix.

## 9.5 Analysis of Case Study

The application of the reliability measures introduced in Section 9.3.2 is demonstrated through the analysis of the two datasets described in Sections 2.3 and 2.4. Using various datasets that are measuring different attributes brings out just how broadly these measures can be used to assess reliability of measuring tools across different fields of research. Both 1PL and 2PL models were fitted using the NLMIXED procedure in SAS, which employs adaptive Gauss-Hermite quadrature to compute the integrals and ultimately the maximum likelihood based parameter estimates.

Results for the LSAT6 data are presented in Table 9.3. These generally indicate low reliability both at the expected item score level and at the expected sum score level. With a five-item measuring tool, this is not surprising as these may not be enough to capture all the relevant information. In addition, the negative estimates

Table 9.1: *Expected sum score reliability estimated for 1PL and 2PL models, using various approximation methods.* $\rho_{S_A}$ *gives the reliability estimated using the Taylor series approximation,* $\rho_f$*, uses the Fisher information measure,* $\rho_{S_l}$ *is the latent variance of logistic regression, i.e.,* $\frac{\pi^2}{3}$*, is used as variance of error* ($\sigma^2_\varepsilon$) *and* $\rho_S$ *is the exact sum score reliability.* $\sigma^2_\theta$ *is the variance of person trait.*

| | | Measure of Reliability | | | |
|---|---|---|---|---|---|
| $\sigma^2_\theta$ | Model | $\rho_{S_A}$ | $\rho_f$ | $\rho_{S_l}$ | $\rho_S$ |
| 0.25 | 1PL | 0.461 | -0.120 | 0.646 | 0.480 |
| | 2PL | 0.687 | 0.615 | 0.889 | 0.704 |
| 1 | 1PL | 0.774 | 0.676 | 0.879 | 0.778 |
| | 2PL | 0.898 | 0.881 | 0.970 | 0.895 |
| 4 | 1PL | 0.932 | 0.887 | 0.967 | 0.928 |
| | 2PL | 0.972 | 0.891 | 0.992 | 0.963 |

for the difficulty parameters for both the 1PL and 2PL models indicate a low difficulty level for all items and the small person trait variance of 0.570 suggests that examinees exhibit similar levels of ability, a scenario that is well known to be less informative. Latent based reliability is larger than the Taylor series approximation and in this case this would lead to different conclusions regarding reliability of the expected sum score. The Fisher information based measure is not useful in this case given that it is negative, which can also be attributed to the relatively small number of items.

The questionnaire for Verbal Aggression data has 24 items, which can be considered of average length and according to Table 9.4, the expected sum score reliability is high: above 0.85 for all measures. Further, the item difficulty estimates for 1PL and 2PL models have maxima, 2.976 and 2.439, and minima, $-1.748$ and $-1.212$, respectively, suggesting a mixture of high and low difficulty level items. The person trait variance of 1.919 suggests a cross-section of persons with varying abilities, forming a desirable scenario to achieve high reliability. Again, the latent correlation based reliability is higher than the manifest correlation based counterpart, although in this case, similar conclusions regarding reliability of both the expected item and the sum score would be reached, regardless of the measure of reliability used.

Table 9.2: *Item reliability for 1PL and 2PL, where person trait variance* $\sigma_\theta^2 = 1$. $\rho_{S_A}$ *is the item reliability obtained using the Taylor series approximation,* $\rho_{S_l}$*, is the latent correlation based item reliability and* $\rho_S$ *is the exact item reliability and finally,* $\beta$ *and* $\alpha$ *are the simulated item difficulty and discrimination parameter values.*

| | 1PL | | | | 2PL | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $\beta$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ | $\beta$ | $\alpha$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ |
| 1 | 1.221 | 0.233 | 0.150 | 0.168 | 0.444 | 3.441 | 0.782 | 0.634 | 0.608 |
| 2 | -0.670 | 0.233 | 0.183 | 0.173 | 3.785 | 1.210 | 0.308 | 0.015 | 0.054 |
| 3 | 1.205 | 0.233 | 0.151 | 0.169 | 0.974 | 3.731 | 0.809 | 0.258 | 0.608 |
| 4 | 0.162 | 0.233 | 0.199 | 0.185 | -1.389 | 2.233 | 0.602 | 0.170 | 0.352 |
| 5 | 0.780 | 0.233 | 0.177 | 0.179 | -0.490 | 2.043 | 0.559 | 0.450 | 0.398 |
| 6 | -3.850 | 0.233 | 0.020 | 0.047 | 3.008 | 1.789 | 0.493 | 0.014 | 0.135 |
| 7 | 3.330 | 0.233 | 0.032 | 0.071 | -0.623 | 2.680 | 0.686 | 0.489 | 0.487 |
| 8 | -0.197 | 0.233 | 0.198 | 0.182 | 2.971 | 1.960 | 0.538 | 0.011 | 0.150 |
| 9 | 2.373 | 0.233 | 0.072 | 0.119 | 1.244 | 2.602 | 0.673 | 0.197 | 0.465 |
| 10 | 2.037 | 0.233 | 0.093 | 0.135 | -0.711 | 2.517 | 0.658 | 0.437 | 0.458 |
| 11 | 0.423 | 0.233 | 0.193 | 0.184 | -1.353 | 2.350 | 0.627 | 0.175 | 0.372 |
| 12 | -1.702 | 0.233 | 0.115 | 0.136 | -2.184 | 0.590 | 0.096 | 0.056 | 0.061 |
| 13 | 0.948 | 0.233 | 0.168 | 0.176 | -2.332 | 2.241 | 0.604 | 0.026 | 0.267 |
| 14 | 2.979 | 0.233 | 0.044 | 0.088 | -0.061 | 2.757 | 0.698 | 0.654 | 0.521 |
| 15 | 3.873 | 0.233 | 0.020 | 0.049 | -2.552 | 2.274 | 0.611 | 0.015 | 0.251 |
| 16 | 0.288 | 0.233 | 0.197 | 0.185 | -0.519 | 2.020 | 0.554 | 0.440 | 0.393 |
| 17 | 0.042 | 0.233 | 0.200 | 0.185 | -1.281 | 2.046 | 0.560 | 0.209 | 0.336 |
| 18 | -1.461 | 0.233 | 0.133 | 0.146 | 0.118 | 0.395 | 0.045 | 0.038 | 0.040 |
| 19 | 0.480 | 0.233 | 0.191 | 0.184 | -0.156 | 2.106 | 0.574 | 0.519 | 0.423 |
| 20 | -3.227 | 0.233 | 0.035 | 0.069 | -1.769 | 1.286 | 0.334 | 0.123 | 0.177 |
| 21 | 3.741 | 0.233 | 0.022 | 0.054 | 3.688 | 2.413 | 0.639 | 0.001 | 0.053 |
| 22 | 0.096 | 0.233 | 0.200 | 0.185 | -1.501 | 2.011 | 0.551 | 0.152 | 0.309 |
| 23 | -2.821 | 0.233 | 0.050 | 0.086 | 1.581 | 1.981 | 0.544 | 0.136 | 0.342 |
| 24 | 3.780 | 0.233 | 0.021 | 0.052 | 0.630 | 1.619 | 0.443 | 0.338 | 0.330 |

## 9.6 Discussion

Beyond doubt, reliability measures based on manifest correlations are of considerable importance in IRT. The reason for their relatively rare use can be largely attributed to lack of efficient means of estimation given that marginalizing the joint distribution of normal random effects, combined with binary data distributions is computationally challenging.

This chapter has outlined a procedure for approximating reliability measures based

on manifest correlations, and illustrated their application for 1PL and 2PL models for both expected item score and expected sum score. We have further explored the relationship between latent and manifest correlations based reliability measures, where it was shown that latent based reliability measures are always greater than their manifest correlation based counterparts. Hence, using one in place of the other should be avoided. A simulation study to assess the performance of the newly introduced Taylor series based reliability measures indicated that they give a true reflection of the exact reliability, especially at the expected sum score level. In comparison to Fisher information based sum score reliability, Taylor series based approximations, perform consistently better, including in the cases where the Fisher information based measure gives negative values, which are not meaningful.

Taylor series based reliability measures do not involve evaluation of integrals with no closed forms. Rather, they use quantities that are easily obtained during model estimation; computation can be handled by most standard statistical software tools. Thus, they represent a less computationally intensive, readily available solution to obtaining reliability of either item score or sum score, whichever truly reflects the required reliability. However, the quality of reliability estimates heavily relies on the quality of model estimated parameters. For example, results from poorly converged models may not reflect the true reliability.

Generally, our findings are useful and relevant for practice and expand on the available tools for measuring reliability. When studying reliability or generalizability, manifest correlation is a more intuitive measure, as it captures the correlation between what is actually observed, and not what happens at the level of a latent construct.

Table 9.3: *Results from the analysis of the LSAT6 data.* $\widehat{\beta}$ *and* $\widehat{\alpha}$ *are item difficulty and discrimination parameters estimates;* $\rho_{i_A}$ *indicates the Taylor series approximated expected item score reliability;* $\rho_{i_l}$ *corresponds to the latent correlation based counterpart;* $\rho_s$ *gives the expected sum score reliability of the corresponding item reliability;* $\rho_f$ *is the Fisher information based reliability measure.*

| | 1PL | | | 2PL | | | |
|---|---|---|---|---|---|---|---|
| item | $\widehat{\beta}$ | $\rho_{i_A}$ | $\rho_{i_l}$ | $\widehat{\beta}$ | $\widehat{\alpha}$ | $\rho_{i_A}$ | $\rho_{i_l}$ |
| 1 | -2.730 | 0.032 | 0.148 | -3.359 | 0.826 | 0.036 | 0.172 |
| 2 | -0.999 | 0.101 | 0.148 | -1.370 | 0.723 | 0.094 | 0.137 |
| 3 | -0.240 | 0.123 | 0.148 | -0.280 | 0.891 | 0.163 | 0.194 |
| 4 | -1.306 | 0.087 | 0.148 | -1.866 | 0.688 | 0.074 | 0.126 |
| 5 | -2.099 | 0.053 | 0.148 | -3.126 | 0.657 | 0.042 | 0.116 |
| $\rho_S$ | | 0.304 | 0.464 | | | 0.312 | 0.465 |
| $\rho_f$ | | -1.343 | | | -1.240 | | |

Table 9.4: *Results from the analysis of the Verbal Aggression Data.* $\widehat{\beta}$ *is the estimate of item difficulty;* $\widehat{\alpha}$*, is the discrimination parameter estimate;* $\rho_{i_A}$ *indicates the Taylor series approximated expected item score reliability;* $\rho_{i_l}$ *corresponds to the latent correlation based counterpart;* $\rho_s$ *gives the expected sum score reliability of the corresponding item reliability;* $\rho_f$ *is the Fisher information based reliability measure.*

| | 1PL | | | 2PL | | | |
|---|---|---|---|---|---|---|---|
| item | $\widehat{\beta}$ | $\rho_{i_A}$ | $\rho_{i_l}$ | $\widehat{\beta}$ | $\widehat{\alpha}$ | $\rho_{i_A}$ | $\rho_{i_l}$ |
| 1 | -1.221 | 0.252 | 0.368 | -0.886 | 1.372 | 0.249 | 0.364 |
| 2 | -0.565 | 0.307 | 0.368 | -0.387 | 1.551 | 0.355 | 0.422 |
| 3 | -0.080 | 0.324 | 0.368 | -0.062 | 1.373 | 0.320 | 0.364 |
| 4 | -1.748 | 0.195 | 0.368 | -1.212 | 1.483 | 0.211 | 0.400 |
| 5 | -0.707 | 0.298 | 0.368 | -0.476 | 1.601 | 0.357 | 0.438 |
| 6 | -0.012 | 0.324 | 0.368 | -0.012 | 1.285 | 0.292 | 0.334 |
| 7 | -0.529 | 0.309 | 0.368 | -0.510 | 0.891 | 0.159 | 0.194 |
| 8 | 0.686 | 0.299 | 0.368 | 0.479 | 1.436 | 0.315 | 0.385 |
| 9 | 1.527 | 0.220 | 0.368 | 1.438 | 0.933 | 0.125 | 0.209 |
| 10 | -1.082 | 0.266 | 0.368 | -0.877 | 1.148 | 0.205 | 0.286 |
| 11 | 0.349 | 0.318 | 0.368 | 0.223 | 1.628 | 0.391 | 0.446 |
| 12 | 1.044 | 0.270 | 0.368 | 0.935 | 0.996 | 0.167 | 0.232 |
| 13 | -1.221 | 0.252 | 0.368 | -0.786 | 1.720 | 0.326 | 0.473 |
| 14 | -0.389 | 0.316 | 0.368 | -0.230 | 2.351 | 0.563 | 0.627 |
| 15 | 0.871 | 0.285 | 0.368 | 0.606 | 1.451 | 0.304 | 0.390 |
| 16 | -0.872 | 0.285 | 0.368 | -0.602 | 1.512 | 0.319 | 0.410 |
| 17 | 0.057 | 0.324 | 0.368 | 0.023 | 2.030 | 0.507 | 0.556 |
| 18 | 1.482 | 0.225 | 0.368 | 0.963 | 1.656 | 0.278 | 0.454 |
| 19 | 0.211 | 0.322 | 0.368 | 0.173 | 1.116 | 0.236 | 0.274 |
| 20 | 1.504 | 0.222 | 0.368 | 1.094 | 1.361 | 0.218 | 0.360 |
| 21 | 2.976 | 0.081 | 0.368 | 2.439 | 1.140 | 0.067 | 0.283 |
| 22 | -0.707 | 0.298 | 0.368 | -0.510 | 1.401 | 0.302 | 0.373 |
| 23 | 0.384 | 0.316 | 0.368 | 0.261 | 1.471 | 0.343 | 0.397 |
| 24 | 2.000 | 0.168 | 0.368 | 1.571 | 1.209 | 0.142 | 0.307 |
| $\rho_S$ | | 0.900 | 0.933 | | | 0.908 | 0.936 |
| $\rho_f$ | | 0.859 | | | 0.863 | | |

# Chapter 10

# Concluding Remarks and Further Research

## 10.1 Concluding Remarks

### 10.1.1 Flexible Methodology For Hierarchical Data and Data with Selection Bias

Motivated by an interesting case study that aims at quantifying expert opinion on clusters of compounds marked for acquisition, we have proposed solutions to statistical problems arising from high-dimensional nature of the data and bias originating from less restricted assignment of the clusters to the experts.

In our quest to quantify expert opinion on the potential of clusters of chemical compounds marked for acquisition, in Chapter 3 we have presented *permutational-splitting sample procedure*, that involves splitting the data into well calculated sub-samples, maximum likelihood estimation within each sub-sample, permuting and re-splitting the data to improve the estimates, and using Monte Carlo methods to approximate integrals, as a means of overcoming the high-dimension problem. While research in high-dimensional problems is at an advanced level, the new procedure was necessary because, unlike many high-dimensional methods that deal with either variable selection, large sample size, or long repeated response vector, our problem had both, long repeated response vector and high-dimensional fixed-effects vector, hence existing methods needed modification.

Results of a simulation study, assessing the performance of the procedure, in compar-

ison to full maximum likelihood estimation, show that parameter estimates and confidence intervals, from the procedure are similar to those obtained through maximum likelihood estimation using a non-split dataset. Parameter estimates showed minimal bias, coverage of confidence intervals for success probability was slightly higher than the expected 95%, but the range was narrow enough for meaningful inferences. Loss in precision, apparent in wider confidence intervals is anticipated, since the procedure splits the data into dependent sub-samples, resulting into a less efficient random-effect variance estimate, which is used in estimating the confidence intervals.

A key issue to guarantee the validity of the process of evaluating the clusters, is the mechanism used to assign the clusters to the experts. It was shown that disregarding this process when quantifying expert opinion can produce misleading results. Essentially, to guarantee validity of the results while using less complicated techniques that are less prone to error, one needs to ensure that the selection and rating processes are independent or that the selection process does not depend on the expert characteristics and does not share any parameter with the rating process. The random allocation of the clusters to the experts seems to be the most, if not the only, practical way to achieve these conditions. Therefore, we strongly advocate for its use in the present work.

Even in carefully designed studies it is not always possible to avoid bias in the estimates of the parameters of interest, for example, in the case study described in Section 2.1. For practical reasons, experts were unknowingly given the choice to select the number of clusters they want to rate, which introduced selection bias. In Chapter 4, we recommended jointly fitting the complex hierarchical models describing the selection and rating mechanisms in order to obtain valid estimates. However, it has been shown that misspecifying the selection model may introduce severe bias in the estimates of the relevant parameters. In response, we have introduced a new approach using the so-called combined model that accounts for the selection process using a new set of random effects. Simulations results clearly showed that, unlike the naive and joint model approaches, the combined model seems to produce unbiased, although less precise, estimates in most settings. This loss of precision may be seen as the price to pay for the robustness archived by the model. When factors suspected to drive the selection process are known and available, the combined model is still useful as a sensitivity analysis tool.

As noted in Chapter 4, blindly assuming that unobserved heterogeneity in repeated measurements data only comes from the correlation in the responses can be too restrictive, as other unobserved sources of variation, like bias also need to be taken into account. In Chapter 6, we have shown through simulations that ignoring other

sources of variation in hierarchical model can have dire consequences on estimation of some covariate effects and their standard errors, as well as, on the variance components and the Type I error rates. Though we used a specific source of variation, namely, overdispersion in count data, it is obvious that the results are applicable to other sources of variation, like bias studied in Chapters 4 and 5. Importantly, we found that the Type I error rates were considerably inflated when overdispersion was ignored, implying that the probability of detecting a spurious effect increases. Remarkably, our findings are strikingly similar to those reported by Litière, Alonso, and Molenberghs (2008) when studying the impact of misspecifying the random effects distribution in a logistic model with a random intercept. It is interesting to see that two related but different types of misspecification, i.e., ignoring overdispersion and misspecifying the random effect distribution, may have very similar consequences.

### 10.1.2 Flexible Methodology For Data With Random Sample Size

In Chapter 7 we have considered the consequences for statistical inference of a random sample size. Our setting is that of univariate random variables from the exponential family that are subject to a stopping rule such that the sample size is either $N = n$ or $N = 2n$, with $n$ specified by design. The stopping rule is stochastic and is allowed to depend on the sample sum $K$ over the first $n$ observations. The rule is generic in the sense that its limiting cases are a deterministic stopping rule, such as in a sequential trial, and a completely random sample size, independent of the data. This setting extends those of both Liu *et al.* (2006) and Molenberghs *et al.* (2013); the former restrict attention to a deterministic stopping rule, although they do so for an arbitrary number of interim looks. The latter confined attention to normally distributed outcomes only.

We have focused on three important inferential aspects. First, we have shown that the sufficient statistic $(K, N)$ is incomplete. Second, we have examined the consequences of this for the sample average, as well as for linear generalizations thereof. We have shown that there is small-sample bias, except for the CRSS case. Even then, there is no optimal estimator, except for the exponential distribution, for which the optimum differs from the ordinary sample average. Third, we have studied maximum likelihood estimation in both a joint as well as a conditional framework. The joint likelihood is for the exponential-family parameter and the stopping rule simultaneously. The conditional likelihood starts from the conditional distribution of the outcomes, given the sample size. Also here, counterintuitive results are derived. The

joint likelihood produces the sample average as maximum likelihood estimator, which is biased in finite samples but is asymptotically unbiased, provided a regularity condition on the stopping rule applies. The conditional likelihood estimator is unbiased, even in small samples. This notwithstanding, the sample average has smaller MSE than the conditional estimator in many important cases, such as the normal and binary examples considered, as well as when the variance of the outcomes is sufficiently small. Under regularity conditions, both estimators are asymptotically equivalent, with the difference between both being $\mathcal{O}(n^{-1})$. The regularity condition is not very restrictive; it essentially comes down to requiring that $F'(k = n\mu)$ approaches zero where $F$ is the stopping rule. For broad classes of parametric functions, this condition is satisfied. We have shown that the corresponding conditional expectations are unbiased.

Hence, when the regularity conditions are satisfied, the sample average remains an attractive and sensible choice for sequential trials. Thus, while some familiar inferential properties no longer hold, estimation after sequential trials is more straightforward than commonly considered and there is little need for complicated, modified estimators, given that the ordinary sample average is acceptable for wide classes of stopping rules, whether stochastic or deterministic.

Molenberghs *et al.* (2013) considered several ramifications of their developments. They commented on the situation of an arbitrary number of looks in a sequential trial, and considered in detail the CRSS case for more than two possible sample sizes. All of this was done for normally distributed outcomes. They also commented on the connection between their derivations and longitudinal outcomes subject to dropout of an MAR type, where dropout depends on observed but not further on unobserved outcomes. While similar, there are subtle differences because now the randomness in the sample size pertains to the number of measurements per subject, rather than to the number of subjects. The difference lies in the fact that measurements within a subject are not independent. Our results extend to these settings as well for the exponential family. Furthermore, connections can be made with a variety of other settings with random sample sizes, such as clustered data with informative cluster sizes, time-to-event data subject to censoring, jointly observed longitudinal and time-to-event data, and random observation times. These settings are currently scrutinized further, and will be reported in a separate manuscript.

While the settings discussed above are enlightening and intuitive, in practice, more than two looks are commonly required. We therefore extended the methodology to general number of looks with the GST framework in Chapter 8, since it is the most used form of RSS trials. Our findings can be summarized as follows.

First, although for a sequential trial with a deterministic stopping rule the ordinary sample average is finite sample biased, it can be been shown both directly and through likelihood arguments that it is asymptotically unbiased and so remains a good candidate for practical use. Further, it is computationally trivial, has a correspondingly simple estimator of precision, derived from observed information matrix and hence a well behaved asymptotic likelihood-based confidence interval. In addition, the mean square error of the sample average is smaller than that of the estimator based on the conditional likelihood, even though the latter is finite sample unbiased. The conditional estimator is also computationally more involved, because there is no closed-form solution. Asymptotically, the mean square errors of both estimators converge to each other.

Second, there is the subtle issue that the sample average may be asymptotically biased for certain stopping rules, when its expectation is considered conditionally on certain values of the sample size. However, this is not a real practical problem because this occurs only for sample sizes that have asymptotic probability zero of being realized.

In addition to estimation in trials with RSS,in Chapter 9 we have also looked at the aspect of reliability of a scale of measurement in studies that measure some unobserved traits, like quality of life, intelligence and some specific behaviors. Based on manifest correlation functions, reliability for a continuous scale is easily obtained as the ratio of true over observed variance, since manifest and latent correlations coincide, a phenomenon that is not carried forward to the categorical scale. Although reliability on a latent scale is easily obtained for the categorical scale, it is usually not of interest. Rather, reliability measures based on manifest correlations are of considerable importance. The reason for their relatively rare use is to a large extent the lack of efficient means of estimation given that marginalizing the joint distribution of normal random effects, combined with binary data distributions is computationally challenging.

We have outlined a procedure for approximating reliability measures based on manifest correlations, and illustrated their application for 1PL and 2PL models for both expected item score and expected sum score. We have further explored the relationship between latent and manifest correlations based reliability measures, where it was shown that latent based reliability measures are always greater than their manifest correlation based counterparts. Hence, using one in place of the other should be avoided. A simulation study to assess the performance of the newly introduced Taylor series based reliability measures indicated that they give a true reflection of the exact reliability, especially at the expected sum score level. In comparison to Fisher

information based sum score reliability, Taylor series based approximations, perform consistently better, including in the cases where the Fisher information based measure gives negative values, which are not meaningful.

Taylor series based reliability measures do not involve evaluation of integrals without closed forms. Rather, they use quantities that are easily obtained during model estimation; computation can be handled by most standard statistical software tools. Thus, they represent a less computationally intensive, readily available solution to obtaining reliability of either item score or sum score, whichever truly reflects the required reliability. However, the quality of reliability estimates heavily relies on the quality of model estimated parameters. For example, results from poorly converged models may not reflect the true reliability.

Generally, our findings are useful and relevant for practice and expand on the available tools for measuring reliability. When studying reliability or generalizability, manifest correlation is a more intuitive measure, as it captures the correlation between what is actually observed, and not what happens at the level of a latent construct.

## 10.2 Further Research

### 10.2.1 Connections Between Combined Model and Missing Data Methodology

Given the robust performance of the combined model in the presence of bias, and given the relationship between bias and missingness, it is only logical that we explore possible connections between the combined model and models for missing data. There is potential that it can lead to a powerful modeling framework for missing data problems.

### 10.2.2 Reliability Measures for Models Multidimensional Traits

Reliability for measurement scales assuming multidimensional traits is still a gray area in measurement studies. The advantage of the reliability approximation functions that we have presented is that they can be easily extended to such settings. Such an extension will be a valuable addition to reliability measures research.

# Bibliography


Agrafiotis, D. K., Alex, S., Dai, H., Derkinderen, A., Farnum, M., Gates, P., Izrailev, S., Jaeger, E. P., Konstant, P., Leung, A., Lobanov,V. S., Marichal, P., Martin, D., Rassokhin, D. N., Shemanarev, M., Skalkin, A., Stong, J., Tabruyn, T., Vermeiren, M., Wan, J., Xu, X. Y., and Yao, X. (2007). Advanced Biological and Chemical Discovery (ABCD): Centralizing discovery knowledge in an inherently decentralized world. *J. Chem. Inf. Model*, **47**, 1999-2014.

Agresti, A. (2002). *Categorical Data Analysis* (2nd ed.). New York: John Wiley & Sons.

Agresti, A., Caffo, B., and Ohman-Strickland, P. (2004). Examples in which misspecifcation of a random effects distribution reduces efficiency, and possible remedies. *Computational Statistics & Data Analysis*, **47**, 639–653.

Alonso, A. and Molenberghs, G. (2008). Surrogate endpoints: Hopes and perils. *Pharmacoeconomics and Outcomes Research*, , 255–259.

Alonso, A., Milanzi, E., Molenberghs, G., Buyck, C., and Bijnens, L. (2013). Impact of selection bias on the qualitative assessment of Cluster of Chemical compounds. *Submitted for publication and in Revision.*

Armitage, P. (1975) *Sequential Medical Trials.* Oxford: Blackwell.

Barndorff-Nielsen, O. and Cox, D.R. (1984). The effect of sampling rules on likelihood statistics. *International Statistical Review*, **52**, 309–326.

Baser, O., Bradley, C.J., Gardiner, J. C., and Given, C. (2003). Testing and correcting for mon-random selection bias due to censoring: An application to medical costs. *Health Services & Outcomes Research Methodology*, **4**, 93–107.

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, Q., Heels-Ansedell, D., Walter, S.D., Guyatt, G.H. and the STOPIT-2 Study Group. (2010). Stopping randomized trials early for benefit and estimation of treatment effects. Systematic review and meta-regression analysis. *Journal of the American Medical Association*,**303**, 1180–1187.

Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhya*, **15**, 377–380.

Berger, M. and Wong, W.(2009) *An introduction to optimal designs for social and biomedical research*. Oxford : Wiley-Blackwell.

Birkel, T. (1992). Laws of large numbers under dependence assumptions. *Statistics & Probability Letters*, **14**, 355–362.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, **18** ,105-110.

Bock, R.D., and Lieberman M. (1970). Fitting a response curve model for dichotomously scored items. *Psychometrika*, **35**, 179–198.

Bond, T.G. & Fox, C.M. (2007) . *Applying the Rasch Model: Fundamental measurement in the human sciences (2nd ed.)*. New Jersey: Lawrence Erlbaum Associates.

Booth, J.G., Casella, G., Friedl, H., and Hobert, J.P. (2003). Negative binomial loglinear mixed models. *Statistical Modelling*, **3**:179–181.

Briggs, D.C., and Wilson, M. (2007). Generalizability in item response modelling. *Journal of Education Measurement*, **44**, 131–155.

Casella, G. and Berger, R.L. (2001). *Statistical Inference*. Pacific Grove: Duxbury Press.

Chen, X. and Xie, M. (2012). A split-and-conquer approach for analysis of extra ordinary large data. *DIMACS technical report 2012-01*, [cited 2013 June 15], **Available from:**
http://dimacs.rutgers.edu/TechnicalReports/TechReports/2012/2012-01.pdf

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman & Hall.

Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., and Kenward, M.G. (2011). Generalized shared-parameter models and missingness at random. *Statistical Modeling*, **11**, 279–311.

Cronbach, L.J., and Shavelson, R.J. (2004). My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement*, **64**, 391-418.

Culligan, B. (2008). Estimating word difficulty using Yes/No tests in an IRT framework and its application for pedagogic objectives. Unpublished dissertation. Tokyo: Temple University Japan.

Davidian, M. and Giltinan, D.M (1995) *Nonlinear Models for Repeated Measurement Data*. New York: Chapman & Hall.

De Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York, NY: The Guilford Press.

De Boeck, P. and Wilson, M. (2004).(Eds.) *Explanatory Item Response Models: A Generalized Linear and non-linear Approach*. New York: Springer.

De Boeck, P. (2009). Random Item Item Response Theory Models. *Technical Report*.

Diggle, P.J. and Kenward, M.G. (1994). Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49-93.

Dimitrov, D.M. (2003). Marginal true score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measuremen*, **27**, 440–458.

Donoho, D.L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *Aide-Memoire*, [cited 2013 June 15], **Available from:**http://www-stat.stanford.edu/ donoho/Lectures/AMS2000/Curses.pdf

Doran, H., Bates, D., Bliese, P., and Dowling, M. (2007). Estimating the multilevel Rasch model with the lme4 package. *Journal of Statistical Software*, **20**, 1–18.

Duchateau, L. and Janssen, P. (2007). *The Frailty Model*. New York: Springer.

Dunbar, J. B. (2000). Compound acquisition strategies. *Pacific Symposium on Biocomputing*, **5**, 552–562.

Emerson,S.S., Gillen, D.L., Kittelson, J.K., Emerson, S.C., and Levin, G.P (2012). RCTdesign: Group Sequential Trial Design. R package version 1.0.

Emerson, S.S. and Fleming, T.R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, **77**, 875–892.

Emerson, S.S. (1988). Parameter estimation following group sequential hypothesis testing. *PhD dissertation*. University of Washington.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, **32**, 928–961.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J.R Statist. Soc. B*, **74**, 37–65.

Faught, E., Wilder, B.J., Ramsay, R.E., Reife, R.A., Kramer, L.D.,Pledger, G.W., and Karim, R.M. (1990.) Topiramate placebo-controlled dose-ranging trial in refractory partialepilepsy using 200-, 400-, and 600-mg daily dosages. *Neurology*, **46**, 1684–1690.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for joint modelling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, **52**, 399–433.

Follmann, D. and Wu, M. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, **51**, 151–168.

Gaure, S. (2013). OLS with Multiple High Dimensional Category Variables. *Computational Statistics and Data Analysis*, **66**, 8-18.

Genelletti, S., Mason, A., and Best, N. (2011). Adjusting for selection effects in epidemiologic studies: Why sensitivity analysis is the only "solution." *Commentary in Epidemiology*, **22**, 36–39.

Geneletti, S., Richardson, S., and Best, N. (2009) Adjusting for selection bias in retrospective, case-control studies. *Biostatistics*, **10**, 17–31.

Grambsch, P. (1983). Sequential sampling based on the observed Fisher information to guarantee the accuracy of the maximum likelihood estimator. *Annals of Statistics*, **11**, 68–77.

Guimaraes, P. and Portugal, P. (2010). A Simple Feasible Alternative Procedure to Estimate Models with High-Dimensional Fixed Effects. *Stata Journal*, **10**, 628–649.

Hack, M.D., Rassokhin, D.N., Buyck, C., Seierstad, M., Skalkin, A., ten Holte, P., Jones, T.K., Mirzadegan, T., Agrafiotis, D.K. (2011). Library enhancement through

the wisdom of crowds. *Journal of Chemical Information and Modeling*, **51**, 3275–3286.

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage Press.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161,

Hernán, M.A., Hernández-Diaz, S., and Robins, J.M. (2004). A structural approach to selection bias. *Epidemiology*, **15**, 615–625.

Hinde, J. and Demétrio, C.G.B. (1998). *Overdispersion: Models and Estimation.* São Paulo: XIII Sinape.

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Horwitz, R. and Feinstein, A. (1978). Alternative analytic methods for case-control studies of estrogens and endometrial cancer. *New England Journal of Medicine*, **299**, 368–387.

Hughes, M.D. and Pocock, S.J. (1988). Stopping rules and estimation problems in clinical trials. *Statistics in Medicine*, **7**, 1231–1242.

Jüni, P. and Egger, M. (2005). Empirical evidence of attrition bias in clinical trials. *International Journal of Epidemiology*, **34**, 87–88.

Jennison, C. and Turnbull, B.W. (2000). *Group Sequential Methods With Applications to Clinical Trials.* London: Chapman & Hall/CRC.

Johnson, N.L. and Kotz, S. (1970). *Distributions in Statistics, Continuous Univariate Distributions, Vol. 2.* Boston: Houghton-Mifflin.

Johnson, R.A., and Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis*(6th ed.). Upper Saddle River, New Jersey: Pearson–Prentice Hall.

Kenward, M.G. and Carpenter J. (2007). Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, **16**, 199–218.

Kenward, M.G. and Molenberghs, G. (1998) Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **13**, 236–247.

Laenen, A. (2008). *Psychometric Validation of Continuous Rating Scales from Complex Data*. Unpublished PhD Thesis. Hasselt University, Belgium.

Lajiness, M. and Watson, I. (2008). Dissimilarity-based approaches to compound acquisition. *Current Opinion in Chemical Biology*, **12**, 366–371.

Lee, B. and Marsh, L.C. (2000). Sample selection bias correction for missing response observations. *Oxford Bulletin of Economics and Statistics*, **62**, 305–322.

Lehmann, E.L. and Stein, C. (1950). Completeness in the sequential case. *Annals of Mathematical Statistics*,**21**, 376–385.

Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association* **86**, 316–327.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

Lindstrom, M.L. and Bates, D.M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, 1014–1021.

Litière,S., Alonso, A. and Molenberghs, G. (2007). Type I and Type II error under random-effects misspecification in generalized linear mixed models. *Biometrics*, **63**, 1038–1044.

Litière,S., Alonso, A. and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. *Statistics In Medicine* **27**, 3125–3144.

Little, R.J.A., and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, **88**, 125–134.

Little, R.J.A. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, **81**, 471–483.

Little, R.J.A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, **90**, 1112–1121.

Liu, A. and Hall, W.J. (1999). Unbiased estimation following a group sequential test. *Biometrika*, **86**, 71–78.

Liu, A., Hall, W.J., Yu, K.F., and Wu, C. (2006). Estimation following a group sequential test for distributions in the one-parameter exponential family. *Statistica Sinica*, **16**, 165–81.

Lord, F.M. (1980). Unbiased estimators of ability parameters of their variance and of their parallel-forms reliability. *Psychometrika*,**48**, 233–245.

Ma, C., Lazo, J.S., and Xie, X. (2011). Compound Acquisition and Prioritization Algorithm for Constructing Structurally Diverse Compound Libraries. *ACS Combinatorial Science*, **13**, 223–231.

McGraw, K.O., and Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, **1**, 30–46.

Meinshausen, N and Bühlmann, P. (2010). Stability selection.*J.R Statist. Soc. B*, **72**, 417–473.

Mesbah, M., Cole, B.F., and Ting Lee, M.-L. (2002). *Statistical Methods for Quality of Life Studies: Design, Measurement, and Analysis*; Questionnaire reliability under the Rasch model, pp. 155–166. New York: Kluwer Academic Publishers.

Milanzi, E., Alonso, A., Buyck,C., Molenberghs, G. and Bijnens,L. (2013) A permutational-splitting sample procedure to quantify expert opinion on Clusters of chemical compounds using high-dimensional data. *submitted.*

Milanzi, E., Molenberghs, G., Alonso, A., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., and Davidian, M. (2012). Properties of estimators in exponential family settings with observation-based stopping rules. *Submitted for publication.*

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* New York: Springer.

Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies.* New York: Wiley.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007). An extended random-effects approach to modeling repeated, overdispersed count data. *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Beunckens, C., Sotto, C., and Kenward, M.G. (2008) Every missing not at random model has got a missing at random counterpart with equal fit. *Journal of the Royal Statistical Society, Series B*, **70**, 371–388.

Molenberghs, G., Verbeke, G., Demétrio, C., and Vieira, A. (2010). A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science*, **25**:325–347.

Molenberghs, G., Verbeke, G., and Iddi, S. (2011). Pseudo-likelihood methodology for partitioned large and complex samples. *Statistics & Probability Letters*, **81**, 892–901.

Molenberghs, G., Kenward, M.G., Aerts, M., Verbeke, G., Tsiatis, A.A., Davidian, M., Rizopoulos, D. (2012). On random sample size, ignorability, ancillarity, completeness, separability, and degeneracy: sequential trials, random sample sizes, and missing data. *Statistical Methods in Medical Research*, *in Press*.

Nelson, K.P., Lipsitz, S.R., Fitzmaurice, G.M., Ibrahim, J., Parzen, M., and Strawderman, R. (2006). Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics*, **15**, 39–57.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, **79**, 755–762.

Newman, C.M. (1984). Asymptotic independence and limit theorems for positively and negatively dependent random variables (Tong, P.L., Eds). In *Inequalities in Statistics and Probability*(pp 127–140). Hayward,CA: Institute of Mathematical Statistics.

OBrien, R. M. (1995). Generalizability coefficients are reliability coefficients. *Quality and Quantity*, **29**, 421–428.

Oxman, A.D, Lavis, J.N., and Fretheim, A. (2007). Use of evidence in WHO recommendations. *Lancet*, **369**, 1883–1889.

Puhani, P.A. (2000). The Heckman correction for sample selection and its critique. *Journal of Economic Surveys*,**14**, 53–68.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., Tuerlinckx, F., De Boeck, P., and Kuppens P. (2003). A non-linear mixed model framework for item response theory. *Pyschological Methods*, **8**, 185–205.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data, with Applications in R.* Chapman and Hall/CRC: Boca Raton.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, **95**: 63–74.

Rodda, B.E., Tsianco, M.C., Bolognese, J.A., and Kersten, M.K. (1988). Clinical Development (Karl, P.E., Eds). In *Biopharmaceutical statistics for drug development* (pp 21–82). Newyork: Marcel Dekker.

Rodríguez, G. and Elo, I. (2003). Intra-class correlation in random effects models for binary data. *The Stata Journal*, **3**, 32–46.

Rosner, G.L. and Tsiatis, A.A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika*, **75**, 723–729.

Rotnitzky, A. (2009). Inverse probability weighted methods. In: *Longitudinal Data Analysis* (G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs eds.), 453–476. CRC/Chapman & Hall, Boca Raton.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, **66**, 688–701.

Rubin, D.B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Schaeffer, G.A., Carlson, R.E., and Matas, R.L. (1986). Assessing the reliability of criterion-referenced measures used to evaluate health-education programs. *Evaluation Review*, **10**, 115–125.

Schultz, J.R., Ruppel, P.L., and Johnson, M.A. (1988). Pharmaceutical lead discovery and optimization (Karl, P.E., Eds). In *Biopharmaceutical statistics for drug development* (pp 21–82). Newyork: Marcel Dekker.

Siegmund, D. (1978). Estimation following sequential tests. *Biometrika*, **64**, 191–199.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, **15**, 72–101.

Surowiecki, J. (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Boston, New York: Little, Brown & Company.

Ten Have, T. R., and Tran, L. (1999). A comparison of mixed effects logistic regression models for binary response data with two nested levels of clustering. *Statistics In Medicine*, **18**, 947–960.

Todd, S., Whitehead, J., and Facey, K.M. (1996). Point and interval estimation following a sequential clinical trial. *Biometrika*, **83**, 453–461.

Torner, A., Duberg, A., Dickman, P., and Svensson, A. (2010). A proposed method to adjust for selection bias in cohort studies. *American Journal of Epidemiology*, **171**, 602–608.

Tsiatis, A.A., Rosner, G.L., and Mehta, C.R. (1984). Exact confidence intervals following a group sequential test. *Biometrics*, **40**, 797–803.

Van der Linden, W.J. and Hambleton, R.K. (Eds.) (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Van Leeuwen, D.M., Barnes, M.D., and Pase, M. (1998). Generalizability theory: A unified approach to assessing the dependability (reliability) of measurements in the health sciences. *Journal of Outcome Measurement*, **2**, 302–325.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2004). Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Controlled Clinical Trials*, **25**, 13–30.

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., and Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, **61**, 295–304.

Vangeneugden T, Molenberghs G, Laenen A, Geys H, Beunckens C, Sotto C.(2010). Marginal correlation in longitudinal binary data based on generalized linear mixed models. *Communications in Statistics - Theory and Methods*, **39**, 3540–3557.

Vangeneugden, T., Molenberghs, G., Verbeke, G., and Demétrio, C. (2011). Marginal correlation from an extended random-effects model for repeated and overdispersed counts. *Journal of Applied Statistics*, **38**, 215–232.

Vansteelandt, K. (2000). *Formal models for Contextualized Personality Psychology. Unpublished doctoral dissertation*, KU Leuven, Belgium.

Vansteelandt, S., Carpenter, J.R., and Kenward, M.G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators *Methodology*, **6**, 37–48.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

Vonesh, E. F., Green, T., and Schluchter, M. D. (2006). Shared parameter models for the joint analysis of longitudinal data and event times. *Statistics in Medicine*, **25**: 143–163.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, **16**, 117-186.

Whitehead, J. (1997). *The Design and Analysis of Sequential Clinical Trials (2nd ed.).* New York: John Wiley & Sons.

Whitehead, J. (1999). A unified theory for sequential clinical trials. *Statistics in Medicine*, **18**, 2271–2286.

Xia, Y., Tong, H., Li, W.K., and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of Royal Statistical Society Series B*, **64**, 363–410.

Zeger, S.L., Liang, K.-Y., and Albert, P.S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

# Appendix A

# Additional Material Related to the Case Study Revisited

## A.1 Results Emanating From Different Selection Models

The case study was also analyzed with the restrictive model that assumes $\text{corr}(a_i, b_i) = 1$ . Tables A.1 display the results.

Table A.1: *Results from shared parameter models estimated by Laplace. The rating process model is Logit*$[P(Y_{ij}=1|b_i)]=\beta_j+b_i$. *The selection model Logit*$[P(X_{ij}=1|b_i)]=g(.)$ *where* $g(.)$ *is given under the columns "selection models" and the logistic-normal model(Naive) is included for comparison. CID is the cluster identification,* $\widehat{\beta}$ *is the cluster effect estimate,* $\widehat{P}$ *is the estimated success probability and R is its corresponding rank.*

| | | | | Selection Models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Naive | | | $\alpha_j+b_i$ | | | $\beta_j+b_i$ | | | $\alpha+b_i$ | | |
| CID | $\widehat{\beta}$ | $\widehat{P}$ | R | $\widehat{\beta}$ | $\widehat{P}$ | R | $\widehat{\beta}$ | $\widehat{P}$ | R | $\widehat{\beta}$ | $\widehat{P}$ | R |
| 265222 | 2.52 | 0.94 | 1 | -1.10 | 0.34 | 15 | -2.46 | 0.16 | 430 | -1.59 | 0.27 | 62 |
| 295061 | 3.83 | 0.92 | 2 | -0.58 | 0.42 | 1 | -2.47 | 0.16 | 556 | -0.43 | 0.44 | 1 |
| 359957 | 0.49 | 0.87 | 3 | -1.47 | 0.29 | 47 | -2.68 | 0.14 | 2896 | -1.59 | 0.28 | 56 |
| 69850 | 1.07 | 0.82 | 4 | -0.91 | 0.37 | 9 | -2.47 | 0.16 | 594 | -1.15 | 0.33 | 17 |
| 84163 | 5.24 | 0.77 | 5 | -2.03 | 0.22 | 239 | -3.47 | 0.08 | 20619 | -1.97 | 0.23 | 202 |
| 296443 | 2.59 | 0.76 | 6 | -2.31 | 0.19 | 488 | -3.22 | 0.10 | 17009 | -2.54 | 0.17 | 894 |
| 7451 | 1.28 | 0.74 | 7 | -2.06 | 0.22 | 272 | -2.78 | 0.13 | 5050 | -2.06 | 0.21 | 276 |
| 277619 | 1.65 | 0.73 | 8 | -1.97 | 0.23 | 191 | -2.76 | 0.13 | 4286 | -1.77 | 0.25 | 117 |
| 315928 | 2.04 | 0.72 | 9 | -0.73 | 0.39 | 5 | -2.28 | 0.18 | 94 | -0.57 | 0.42 | 3 |
| 296535 | 2.77 | 0.71 | 10 | -0.65 | 0.40 | 3 | -2.71 | 0.14 | 3393 | -0.63 | 0.40 | 4 |
| 313914 | 2.18 | 0.70 | 11 | -1.91 | 0.23 | 166 | -3.33 | 0.09 | 18903 | -1.47 | 0.29 | 43 |
| 277774 | 2.20 | 0.69 | 12 | -1.11 | 0.34 | 18 | -2.63 | 0.14 | 2260 | -1.42 | 0.30 | 40 |
| 178994 | 1.85 | 0.68 | 13 | -0.63 | 0.41 | 2 | -1.98 | 0.21 | 2 | -0.65 | 0.40 | 5 |
| 296560 | 1.89 | 0.66 | 14 | -1.07 | 0.34 | 17 | -2.45 | 0.16 | 408 | -1.00 | 0.35 | 14 |
| 464822 | 1.21 | 0.66 | 15 | -1.45 | 0.29 | 45 | -2.70 | 0.14 | 3057 | -1.48 | 0.29 | 47 |
| 265441 | 1.87 | 0.65 | 16 | -1.95 | 0.23 | 175 | -3.12 | 0.10 | 14099 | -1.98 | 0.22 | 215 |
| 292805 | 1.47 | 0.65 | 17 | -0.72 | 0.39 | 4 | -2.54 | 0.15 | 951 | -0.72 | 0.40 | 6 |
| 432169 | 1.45 | 0.64 | 18 | -1.09 | 0.34 | 20 | -2.89 | 0.12 | 7546 | -1.19 | 0.32 | 24 |
| 292579 | 1.85 | 0.64 | 19 | -1.48 | 0.29 | 48 | -2.79 | 0.13 | 4832 | -1.30 | 0.31 | 28 |
| 278927 | 1.30 | 0.63 | 20 | -1.61 | 0.27 | 59 | -2.41 | 0.16 | 322 | -1.30 | 0.31 | 32 |
| $\widehat{\sigma}^2$ | 20.02 | | | 4.05 | | | 3.06 | | | 3.96 | | |

# Appendix B

# Additional Material Derivation of Properties of Estimators

## B.1 Stopping Probability for Normally Distributed Outcomes

Molenberghs *et al* (2012) derived (7.22) from first principles. It is also possible to derive this expression from (7.9), by plugging in the standard normal density for $\widetilde{f}_1(z)$ with $A(k) = \alpha + \beta k/n$, combined with $f_{n,\theta=\mu} = \varphi_{\mu,n}(k)$.

With these choices, we find:

$$f(N=n) = \exp\left(-\frac{n}{2}\mu^2\right) \cdot I$$

with

$$\begin{aligned} I &= \int_{k=-\infty}^{k=+\infty} \frac{e^{\mu k - \frac{1}{2}\frac{k^2}{n}}}{\sqrt{n}\sqrt{2\pi}} \int_{z=-\infty}^{z=\alpha+\beta k/n} \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} dz\, dk \qquad \text{(B.1)} \\ &= \frac{1}{\sqrt{n}(\sqrt{2\pi})^2} \int_{k=-\infty}^{k=+\infty} \int_{t=-\infty}^{t=\alpha} \exp\left[\mu k - \frac{1}{2}k^2 n - \frac{1}{2}\left(t + \beta k/n\right)^2\right] dt\, dk \\ &= \frac{1}{\sqrt{n}(\sqrt{2\pi})^2} \int_{k=-\infty}^{k=+\infty} \int_{t=-\infty}^{t=\alpha} \exp\left[-\frac{1}{2}\frac{(k-\ell)^2}{q^2} - \frac{1}{2}\frac{(t-L)^2}{Q^2} - \frac{1}{2}M\right] dt\, dk, \end{aligned}$$

with

$$\begin{aligned} \ell &= \frac{n\mu - \beta t}{\beta^2 + n} \cdot n, \\ q &= \frac{n}{\sqrt{\beta^2 + n}}, \\ L &= -\mu\beta, \\ Q &= \frac{\sqrt{\beta^2 + n}}{\sqrt{n}}, \\ M &= -\mu^2 n. \end{aligned}$$

This implies that

$$f(N = n) = \exp\left(-\frac{n}{2}\mu^2\right) \cdot \frac{1}{\sqrt{n}} \cdot q \cdot Q \cdot \Phi\left(\frac{\alpha - L}{Q}\right) \exp\left(+\frac{n}{2}\mu^2\right) = \Phi\left(\frac{\alpha + \beta\mu}{\sqrt{1 + \beta^2/n}}\right),$$

which equals (7.22).

## B.2 Joint Probability for Binary Outcome with Beta-driven Stopping Rule

Considering (7.44), expression (7.48) is valid if we can show that

$$\frac{2n - k}{2n} \left( \begin{array}{c} 2n \\ k \end{array} \right) = \left( \begin{array}{c} 2n \\ k \end{array} \right) - H(k), \tag{B.2}$$

with

$$H(k) = \sum_{z=0\vee(k-n)}^{k\wedge n} \left( \begin{array}{c} n \\ z \end{array} \right) \left( \begin{array}{c} n \\ k - z \end{array} \right) \frac{z}{n}. \tag{B.3}$$

In other words, we need to show

$$\frac{k}{2n} \left( \begin{array}{c} 2n \\ k \end{array} \right) = H(k),$$

i.e.,

$$\begin{aligned} \left( \begin{array}{c} 2n - 1 \\ k - 1 \end{array} \right) &= \sum_{z=1\vee(k-n)}^{k\wedge n} \left( \begin{array}{c} n - 1 \\ z - 1 \end{array} \right) \left( \begin{array}{c} n \\ k - z \end{array} \right) \\ &= \sum_{z'=0\vee(k-1-n)}^{k-1\wedge n-1} \left( \begin{array}{c} n - 1 \\ z' \end{array} \right) \left( \begin{array}{c} n \\ k - 1 - z' \end{array} \right), \end{aligned}$$

which is obviously correct.

## B.3 Conditional Expectations for the Conditional Likelihood in the Binary Case

As stated in Section 7.5.3, it is easy to show from first principles and instructive that the conditional expectations of the conditional scores (7.134) and (7.137).

An explicit expression for $A'_n(\pi)$ in (7.134) is

$$\begin{aligned} A'_n(\pi) &= \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k-1}(1-\pi)^{n-k}F(k) - \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) n\pi^{k}(1-\pi)^{n-k-1}F(k) \\ &\quad + \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k}(1-\pi)^{n-k-1}F(k). \end{aligned}$$

The expectation becomes

$$\begin{aligned} E[S_n(\pi)|N=n] &= E\left[\frac{k}{\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} - \frac{A'_n(\pi)}{A_n(\pi)}\right] \\ &= \frac{\sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k}(1-\pi)^{n-k}F(k)}{A_n(\pi)\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} \\ &\quad - \frac{\sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k}(1-\pi)^{n-k}F(k)}{A_n(\pi)\pi(1-\pi)} \cdot (1-\pi) \\ &\quad + \frac{n\sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) \pi^{k}(1-\pi)^{n-k}F(k)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\ &\quad - \frac{\sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k}(1-\pi)^{n-k}F(k)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\ &= \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} - \frac{n\pi}{\pi(1-\pi)} - \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} \cdot (1-\pi) \\ &\quad + \frac{n}{\pi(1-\pi)} \cdot \pi - \frac{P(\pi)}{A_n(\pi)\pi(1-\pi)} \cdot \pi \\ &= 0, \end{aligned}$$

where

$$P(\pi) = \sum_{k=0}^{n} \left( \begin{array}{c} n \\ k \end{array} \right) k\pi^{k}(1-\pi)^{n-k}F(k)$$

Because (7.137) is slightly more involved than (7.134), we proceed differently.

Writing

$$S_{2n}(\pi) = \frac{1}{\pi(1-\pi)}\left\{K - \frac{2n\pi}{1-A(K,\pi)} + \frac{\sum_{\ell=0}^{2n}\ell\pi^{\ell}(1-\pi)^{2n-\ell}\sum_{z}\left(\begin{array}{c} n \\ z \end{array}\right)\left(\begin{array}{c} n \\ \ell - z \end{array}\right)F(z)}{1-A(\pi)}\right\},$$

with

$$A(\pi) = \sum_{\ell=0}^{2n}\pi^{\ell}(1-\pi)^{2n-\ell}\sum_{z}\left(\begin{array}{c} n \\ z \end{array}\right)\left(\begin{array}{c} n \\ \ell - z \end{array}\right)F(z),$$

and using the result that

$$2n\pi = \sum_{k=0}^{2n} k\pi^{k}(1-\pi)^{2n-k},$$

we can write the score as:

$$\begin{aligned} S_{2n}(\pi) &= \frac{1}{\pi(1-\pi)}\left\{K - \frac{\sum_{\ell=0}^{2n}Kp(\ell,2n)}{1-A(K,\pi)}\right\} \\ &= \frac{1}{\pi(1-\pi)}\left\{K - \frac{\sum_{\ell=0}^{2n}Kp(\ell,2n)}{p(N=2n)}\right\} \end{aligned}$$

which immediately implies that $E[S_{2n}(\pi)|N=2n] = 0$.

# Appendix C

# Additional Simulation Studies for GST

## C.1 Simulation Study for Stopping Rule $\Phi(\alpha + \beta k)$

### C.1.1 Simulation Settings

The results presented in this section are from the simulation study run with the purpose of investigating the behavior of the joint and conditional likelihood estimators in non-fixed sample size trials. The sample size $N$ can take the values $n$ and $2n$.

Specifically, we generated $Y_i \sim N(\mu, 1)$ $i = 1 \dots n$, from which $F = \Phi(\alpha + \beta k)$ is calculated, with $K = \sum_{i=1}^{n} Y_i$. The decision to stop or continue is reached by generating $Q \sim U(0, 1)$ and that if $Q \leq F$, the trial stops, otherwise we generate another $Y_i \sim N(\mu, 1)$ $i = n + 1 \dots 2n$. Finally the estimate of $\mu$ is obtained by maximizing the relevant likelihood (joint or conditional). The following values were considered: $\mu = 2; 4; 10$ and $n = 25; 50; 250; 500; 5000$. To also allow for small effects to show up, a total of 1 million simulations were done for each setting.

### C.1.2 Simulation Results

The results indicate small biases in all cases, the highest bias value being 0.1%, which comes from the conditional likelihood estimator for $N = 25$ and $\mu = 2$. In general, though, the conditional likelihood estimator shows little or no bias. For the sample average, comparing the overall results with the ones conditional on sample size, reveals that the bias is slightly higher in the conditional estimates than the marginal ones for the small sample size. The asymptotic behavior of bias is in line with theory, given

Table C.1: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples; n is the sample size generated at a particular stage, $\mu$ is the true mean,$\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.9985 | .00077 | 0.02550 | 1.68972 | 2.3072 |
| 50 | 2 | 1.9990 | .00052 | 0.01136 | 1.79192 | 2.2060 |
| 250 | 2 | 2.0001 | .00003 | 0.00200 | 1.91241 | 2.0877 |
| 500 | 2 | 2.0000 | .00002 | 0.00100 | 1.93805 | 2.0620 |
| 5000 | 2 | 2.0000 | .00001 | 0.00010 | 1.98041 | 2.0196 |
| 25 | 4 | 3.9991 | .00024 | 0.02272 | 3.70628 | 4.2918 |
| 50 | 4 | 3.9998 | .00004 | 0.01018 | 3.80238 | 4.1973 |
| 250 | 4 | 4.0001 | .00002 | 0.00200 | 3.91241 | 4.0877 |
| 500 | 4 | 4.0000 | .00001 | 0.00100 | 3.93805 | 4.0620 |
| 5000 | 4 | 4.0000 | .00000 | 0.00010 | 3.98041 | 4.0196 |
| 25 | 10 | 10.0001 | .00001 | 0.02009 | 9.72242 | 10.2779 |
| 50 | 10 | 10.0001 | .00000 | 0.01000 | 9.80405 | 10.1961 |
| 250 | 10 | 10.0001 | .00001 | 0.00200 | 9.91241 | 10.0877 |
| 500 | 10 | 10.0000 | .00000 | 0.00100 | 9.93805 | 10.0620 |
| 5000 | 10 | 10.0000 | .00000 | 0.00010 | 9.98041 | 10.0196 |

that it decreases with increasing sample size. Loss of information in the conditional estimates is noticeable but very small in the settings studied, again in line with theory.

Details are provided in Tables C.1–C.6.

## C.2 Simulation Study for Stopping Rule $\Phi(\alpha+\beta k/n)$

### C.2.1 Simulation Settings

The results presented here are from a simulation study run with the purpose of investigating the behavior of joint and conditional likelihood estimators in non-fixed sample size trials. In contrast to Section C.1.1 the stopping rule is now $F = \Phi\left(\alpha + \beta\frac{k}{n}\right)$. All other settings are as in Section C.1.1.

Table C.2: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples where we stopped at $N = n$; $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias= $|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.98573 | 0.00714 | 0.04009 | 1.59438 | 2.3771 |
| 50 | 2 | 1.95704 | 0.02148 | 0.02062 | 1.68982 | 2.2243 |
| 250 | 2 | 1.90287 | 0.04856 | 0.01343 | 1.77891 | 2.0268 |
| 25 | 4 | 3.97778 | 0.00555 | 0.04022 | 3.58737 | 4.3682 |
| 50 | 4 | 3.85891 | 0.03527 | 0.03438 | 3.62663 | 4.0912 |
| 25 | 10 | 9.97525 | 0.00248 | 0.04061 | 9.58325 | 10.3673 |

### C.2.2 Simulation Results

The results show small biases in all cases, the highest bias value being 0.1% which comes from the conditional likelihood estimate for $N = 25$ and $\mu = 2$, though in general the conditional likelihood estimator shows little or no bias. Comparing the marginal and conditional estimators for the sample average reveals that the bias is slightly higher in the conditional estimators than the marginal ones for the small sample size. Also here, the asymptotic behavior of the bias is in line with that expected from the theoretical developments, as it decreases with increasing sample size. Loss of information in the conditional estimator is discernable but small.

The results are presented in Tables C.7–C.12.

Table C.3: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples where we continued to $N = 2n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 2.0033 | .00166 | 0.02001 | 1.72613 | 2.2805 |
| 50 | 2 | 2.0067 | .00332 | 0.01004 | 1.81065 | 2.2027 |
| 250 | 2 | 2.0001 | .00003 | 0.00200 | 1.91241 | 2.0877 |
| 500 | 2 | 2.0000 | .00002 | 0.00100 | 1.93805 | 2.0620 |
| 5000 | 2 | 2.0000 | .00001 | 0.00010 | 1.98041 | 2.0196 |
| 25 | 4 | 4.0025 | .00061 | 0.02001 | 3.72526 | 4.2796 |
| 50 | 4 | 4.0057 | .00144 | 0.01003 | 3.80974 | 4.2017 |
| 250 | 4 | 4.0001 | .00002 | 0.00200 | 3.91241 | 4.0877 |
| 500 | 4 | 4.0000 | .00001 | 0.00100 | 3.93805 | 4.0620 |
| 5000 | 4 | 4.0000 | .00000 | 0.00010 | 3.98041 | 4.0196 |
| 25 | 10 | 10.0003 | .00002 | 0.02000 | 9.72306 | 10.2774 |
| 50 | 10 | 10.0001 | .00000 | 0.01000 | 9.80405 | 10.1961 |
| 250 | 10 | 10.0001 | .00001 | 0.00200 | 9.91241 | 10.0877 |
| 500 | 10 | 10.0000 | .00000 | 0.00100 | 9.93805 | 10.0620 |
| 5000 | 10 | 10.0000 | .00000 | 0.00010 | 9.98041 | 10.0196 |

Table C.4: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples; $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 2.0069 | 0.00687 | 0.02560 | 1.69777 | 2.3160 |
| 50 | 2 | 2.0195 | 0.01951 | 0.01182 | 1.81170 | 2.2273 |
| 25 | 4 | 4.0196 | 0.01961 | 0.02318 | 3.72629 | 4.3129 |
| 50 | 4 | 4.0423 | 0.04227 | 0.01206 | 3.84389 | 4.2406 |
| 25 | 10 | 10.0527 | 0.05272 | 0.02295 | 9.77448 | 10.3310 |

Table C.5: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples where we stopped at $N = n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.98129 | 0.01871 | 0.04029 | 1.58970 | 2.3729 |
| 50 | 2 | 1.95747 | 0.04253 | 0.02062 | 1.68994 | 2.2250 |
| 25 | 4 | 3.97558 | 0.02442 | 0.04035 | 3.58500 | 4.3662 |
| 50 | 4 | 3.88135 | 0.11865 | 0.02860 | 3.64853 | 4.1142 |
| 25 | 10 | 9.97511 | 0.02489 | 0.04062 | 9.58309 | 10.3671 |

Table C.6: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples where we continued to $N = 2n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 2.0166 | 0.01661 | 0.02035 | 1.73890 | 2.2943 |
| 50 | 2 | 2.0309 | 0.03089 | 0.01104 | 1.83403 | 2.2277 |
| 25 | 4 | 4.0266 | 0.02664 | 0.02080 | 3.74884 | 4.3044 |
| 50 | 4 | 4.0490 | 0.04901 | 0.01250 | 3.85207 | 4.2459 |
| 25 | 10 | 10.0531 | 0.05308 | 0.02289 | 9.77537 | 10.3308 |

Table C.7: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples, n is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.9997 | .00014 | 0.02904 | 1.67065 | 2.3288 |
| 50 | 2 | 2.0001 | .00007 | 0.01452 | 1.76743 | 2.2329 |
| 250 | 2 | 1.9999 | .00005 | 0.00291 | 1.89581 | 2.1040 |
| 500 | 2 | 2.0000 | .00001 | 0.00145 | 1.92638 | 2.0736 |
| 5000 | 2 | 2.0000 | .00001 | 0.00015 | 1.97675 | 2.0233 |
| 25 | 4 | 3.9998 | .00006 | 0.02888 | 3.67160 | 4.3279 |
| 50 | 4 | 4.0001 | .00002 | 0.01444 | 3.76801 | 4.2322 |
| 250 | 4 | 4.0001 | .00002 | 0.00289 | 3.89630 | 4.1038 |
| 500 | 4 | 4.0000 | .00001 | 0.00144 | 3.92663 | 4.0734 |
| 5000 | 4 | 4.0000 | .00000 | 0.00014 | 3.97681 | 4.0232 |
| 25 | 10 | 9.9997 | .00003 | 0.02841 | 9.67427 | 10.3252 |
| 50 | 10 | 10.0001 | .00001 | 0.01421 | 9.76993 | 10.2303 |
| 250 | 10 | 10.0001 | .00001 | 0.00284 | 9.89715 | 10.1030 |
| 500 | 10 | 10.0000 | .00000 | 0.00142 | 9.92723 | 10.0728 |
| 5000 | 10 | 10.0000 | .00000 | 0.00014 | 9.97700 | 10.0230 |

Table C.8: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples where we stopped at $N = n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 2.0005 | .00026 | 0.0200 | 1.72333 | 2.2777 |
| 50 | 2 | 2.0002 | .00009 | 0.0100 | 1.80418 | 2.1962 |
| 250 | 2 | 2.0001 | .00006 | 0.0020 | 1.91246 | 2.0878 |
| 500 | 2 | 2.0001 | .00002 | 0.0010 | 1.93807 | 2.0620 |
| 5000 | 2 | 2.0000 | .00000 | 0.0001 | 1.98041 | 2.0196 |
| 25 | 4 | 4.0003 | .00007 | 0.0200 | 3.72309 | 4.2775 |
| 50 | 4 | 4.0001 | .00003 | 0.0100 | 3.80412 | 4.1961 |
| 250 | 4 | 4.0001 | .00001 | 0.0020 | 3.91240 | 4.0877 |
| 500 | 4 | 4.0000 | .00001 | 0.0010 | 3.93806 | 4.0620 |
| 5000 | 4 | 4.0000 | .00000 | 0.0001 | 3.98041 | 4.0196 |
| 25 | 10 | 10.0004 | .00004 | 0.0200 | 9.72320 | 10.2776 |
| 50 | 10 | 10.0002 | .00002 | 0.0100 | 9.80421 | 10.1962 |
| 250 | 10 | 10.0001 | .00001 | 0.0020 | 9.91240 | 10.0877 |
| 500 | 10 | 10.0000 | .00000 | 0.0010 | 9.93805 | 10.0620 |
| 5000 | 10 | 10.0000 | .00000 | 0.0001 | 9.98041 | 10.0196 |

Table C.9: *Estimates were obtained by maximizing the joint likelihood and averaging was done over estimates from all the simulated samples where we continued to $N = 2n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.9994 | .00031 | 0.03291 | 1.64807 | 2.3507 |
| 50 | 2 | 2.0001 | .00006 | 0.01646 | 1.75168 | 2.2486 |
| 250 | 2 | 1.9998 | .00009 | 0.00330 | 1.88866 | 2.1110 |
| 500 | 2 | 2.0000 | .00002 | 0.00165 | 1.92137 | 2.0786 |
| 5000 | 2 | 2.0000 | .00001 | 0.00016 | 1.97518 | 2.0249 |
| 25 | 4 | 3.9995 | .00012 | 0.03287 | 3.64846 | 4.3506 |
| 50 | 4 | 4.0001 | .00002 | 0.01643 | 3.75188 | 4.2483 |
| 250 | 4 | 4.0001 | .00002 | 0.00328 | 3.88910 | 4.1111 |
| 500 | 4 | 4.0000 | .00000 | 0.00164 | 3.92154 | 4.0785 |
| 5000 | 4 | 4.0000 | .00000 | 0.00016 | 3.97521 | 4.0248 |
| 25 | 10 | 9.9994 | .00006 | 0.03265 | 9.64959 | 10.3492 |
| 50 | 10 | 10.0001 | .00001 | 0.01634 | 9.75262 | 10.2475 |
| 250 | 10 | 10.0001 | .00001 | 0.00327 | 9.88944 | 10.1107 |
| 500 | 10 | 10.0000 | .00000 | 0.00163 | 9.92178 | 10.0783 |
| 5000 | 10 | 10.0000 | .00000 | 0.00016 | 9.97529 | 10.0248 |

Table C.10: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples; n is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.9997 | .00031 | 0.02904 | 1.67061 | 2.3288 |
| 50 | 2 | 2.0001 | .00010 | 0.01454 | 1.76728 | 2.2329 |
| 250 | 2 | 2.0000 | .00000 | 0.00290 | 1.89594 | 2.1041 |
| 500 | 2 | 2.0001 | .00006 | 0.00713 | 1.85669 | 2.1434 |
| 5000 | 2 | 2.0000 | .00001 | 0.00015 | 1.97675 | 2.0233 |
| 25 | 4 | 4.0000 | .00004 | 0.02888 | 3.67186 | 4.3282 |
| 50 | 4 | 4.0001 | .00013 | 0.01445 | 3.76798 | 4.2323 |
| 250 | 4 | 4.0000 | .00000 | 0.00289 | 3.89622 | 4.1038 |
| 500 | 4 | 4.0001 | .00005 | 0.00144 | 3.92666 | 4.0734 |
| 5000 | 4 | 4.0000 | .00001 | 0.00014 | 3.97681 | 4.0232 |
| 25 | 10 | 10.0000 | .00004 | 0.02841 | 9.67458 | 10.3255 |
| 50 | 10 | 10.0001 | .00010 | 0.01422 | 9.76984 | 10.2304 |
| 250 | 10 | 10.0000 | .00002 | 0.00284 | 9.89709 | 10.1029 |
| 500 | 10 | 10.0001 | .00005 | 0.00142 | 9.92726 | 10.0728 |
| 5000 | 10 | 10.0000 | .00001 | 0.00014 | 9.97700 | 10.0230 |

Table C.11: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples where we stopped at $N = n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 1.9988 | .00118 | 0.04000 | 1.60682 | 2.3908 |
| 50 | 2 | 1.9999 | .00006 | 0.02000 | 1.72276 | 2.2771 |
| 250 | 2 | 1.9999 | .00007 | 0.00400 | 1.87597 | 2.1239 |
| 500 | 2 | 2.0000 | .00002 | 0.01001 | 1.82802 | 2.1720 |
| 5000 | 2 | 2.0000 | .00003 | 0.00020 | 1.97232 | 2.0278 |
| 25 | 4 | 3.9995 | .00055 | 0.04000 | 3.60745 | 4.3915 |
| 50 | 4 | 4.0002 | .00017 | 0.02000 | 3.72298 | 4.2774 |
| 250 | 4 | 3.9998 | .00016 | 0.00400 | 3.87588 | 4.1238 |
| 500 | 4 | 4.0000 | .00003 | 0.00200 | 3.91238 | 4.0877 |
| 5000 | 4 | 4.0000 | .00003 | 0.00020 | 3.97231 | 4.0278 |
| 25 | 10 | 9.9993 | .00068 | 0.04000 | 9.60732 | 10.3913 |
| 50 | 10 | 10.0001 | .00012 | 0.02000 | 9.72293 | 10.2773 |
| 250 | 10 | 9.9999 | .00014 | 0.00400 | 9.87590 | 10.1238 |
| 500 | 10 | 10.0000 | .00003 | 0.00200 | 9.91238 | 10.0877 |
| 5000 | 10 | 10.0000 | .00004 | 0.00020 | 9.97232 | 10.0278 |

Table C.12: *Estimates were obtained by maximizing the conditional likelihood and averaging was done over estimates from all the simulated samples where we continued to $N = 2n$, $n$ is the sample size generated at a particular stage, $\mu$ is the true mean, $\widehat{\mu}$ is the average estimated mean, Rel. bias=$|(\mu - \widehat{\mu})|/\mu$, 'MSE' is the mean square error and 'lower' and 'upper' are the lower and upper confidence interval limits respectively, obtained as $\widehat{\mu} \pm 1.96\widehat{\sigma}$.*

| $n$ | $\mu$ | $\widehat{\mu}$ | Rel. bias | MSE | lower | upper |
|---|---|---|---|---|---|---|
| 25 | 2 | 2.0004 | .00041 | 0.02000 | 1.72322 | 2.2776 |
| 50 | 2 | 2.0002 | .00023 | 0.01000 | 1.80423 | 2.1962 |
| 250 | 2 | 2.0001 | .00006 | 0.00200 | 1.91241 | 2.0877 |
| 500 | 2 | 2.0001 | .00009 | 0.00515 | 1.87638 | 2.1238 |
| 5000 | 2 | 2.0000 | .00000 | 0.00010 | 1.98040 | 2.0196 |
| 25 | 4 | 4.0005 | .00051 | 0.02000 | 3.72332 | 4.2777 |
| 50 | 4 | 4.0001 | .00011 | 0.01000 | 3.80411 | 4.1961 |
| 250 | 4 | 4.0001 | .00013 | 0.00200 | 3.91247 | 4.0878 |
| 500 | 4 | 4.0001 | .00006 | 0.00100 | 3.93808 | 4.0620 |
| 5000 | 4 | 4.0000 | .00000 | 0.00010 | 3.98040 | 4.0196 |
| 25 | 10 | 10.0006 | .00057 | 0.02000 | 9.72338 | 10.2778 |
| 50 | 10 | 10.0001 | .00008 | 0.01000 | 9.80408 | 10.1961 |
| 250 | 10 | 10.0001 | .00013 | 0.00200 | 9.91248 | 10.0878 |
| 500 | 10 | 10.0001 | .00006 | 0.00100 | 9.93808 | 10.0620 |
| 5000 | 10 | 10.0000 | .00000 | 0.00010 | 9.98040 | 10.0196 |

# Appendix D

# Additional Results From The Simulation Study on reliability Measures

Table D.1: *Item reliability for 1PL and 2PL, where person trait variance,* $\sigma_\theta^2 = 4$. $\rho_{S_A}$ *is the item reliability obtained using the Taylor series approximation;* $\rho_{S_l}$ *is the latent correlation based item reliability;* $\rho_S$ *is the exact item reliability;* $\beta$ *and* $\alpha$ *are the simulated item difficulty and discrimination parameter values.*

| | 1PL | | | | 2PL | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $\beta$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ | $\beta$ | $\alpha$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ |
| 1 | 1.221 | 0.549 | 0.413 | 0.407 | 0.444 | 3.441 | 0.935 | 0.874 | 0.782 |
| 2 | -0.670 | 0.549 | 0.472 | 0.398 | 3.785 | 1.210 | 0.640 | 0.056 | 0.380 |
| 3 | 1.205 | 0.549 | 0.415 | 0.407 | 0.974 | 3.731 | 0.944 | 0.582 | 0.817 |
| 4 | 0.162 | 0.549 | 0.498 | 0.410 | -1.389 | 2.233 | 0.858 | 0.451 | 0.655 |
| 5 | 0.780 | 0.549 | 0.463 | 0.411 | -0.490 | 2.043 | 0.835 | 0.766 | 0.642 |
| 6 | -3.850 | 0.549 | 0.075 | 0.268 | 3.008 | 1.789 | 0.795 | 0.055 | 0.555 |
| 7 | 3.330 | 0.549 | 0.118 | 0.337 | -0.623 | 2.680 | 0.897 | 0.793 | 0.715 |
| 8 | -0.197 | 0.549 | 0.498 | 0.406 | 2.971 | 1.960 | 0.824 | 0.043 | 0.586 |
| 9 | 2.373 | 0.549 | 0.238 | 0.377 | 1.244 | 2.602 | 0.892 | 0.496 | 0.736 |
| 10 | 2.037 | 0.549 | 0.290 | 0.389 | -0.711 | 2.517 | 0.885 | 0.756 | 0.700 |
| 11 | 0.423 | 0.549 | 0.489 | 0.411 | -1.353 | 2.350 | 0.870 | 0.459 | 0.672 |
| 12 | -1.702 | 0.549 | 0.343 | 0.367 | -2.184 | 0.590 | 0.297 | 0.191 | 0.197 |
| 13 | 0.948 | 0.549 | 0.446 | 0.410 | -2.332 | 2.241 | 0.859 | 0.097 | 0.598 |
| 14 | 2.979 | 0.549 | 0.155 | 0.353 | -0.061 | 2.757 | 0.902 | 0.883 | 0.723 |
| 15 | 3.873 | 0.549 | 0.074 | 0.309 | -2.552 | 2.274 | 0.863 | 0.058 | 0.589 |
| 16 | 0.288 | 0.549 | 0.495 | 0.411 | -0.519 | 2.020 | 0.832 | 0.758 | 0.638 |
| 17 | 0.042 | 0.549 | 0.500 | 0.409 | -1.281 | 2.046 | 0.836 | 0.514 | 0.631 |
| 18 | -1.461 | 0.549 | 0.379 | 0.376 | 0.118 | 0.395 | 0.160 | 0.135 | 0.130 |
| 19 | 0.480 | 0.549 | 0.486 | 0.411 | -0.156 | 2.106 | 0.844 | 0.812 | 0.653 |
| 20 | -3.227 | 0.549 | 0.128 | 0.297 | -1.769 | 1.286 | 0.668 | 0.359 | 0.450 |
| 21 | 3.741 | 0.549 | 0.083 | 0.316 | 3.688 | 2.413 | 0.876 | 0.003 | 0.652 |
| 22 | 0.096 | 0.549 | 0.499 | 0.409 | -1.501 | 2.011 | 0.831 | 0.418 | 0.617 |
| 23 | -2.821 | 0.549 | 0.175 | 0.316 | 1.581 | 1.981 | 0.827 | 0.386 | 0.644 |
| 24 | 3.780 | 0.549 | 0.080 | 0.314 | 0.630 | 1.619 | 0.761 | 0.671 | 0.585 |

Table D.2: *Item reliability for 1PL and 2PL, where person trait variance,* $\sigma_\theta^2 = 0.25$. $\rho_{S_A}$ *is the item reliability obtained using the Taylor series approximation;* $\rho_{S_l}$ *is the latent correlation based item reliability;* $\rho_S$ *is the exact item reliability;* $\beta$ *and* $\alpha$ *are the simulated item difficulty and discrimination parameter values.*

| | 1PL | | | | 2PL | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Item | $\beta$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ | $\beta$ | $\alpha$ | $\rho_{i_l}$ | $\rho_{i_A}$ | $\rho_i$ |
| 1 | 1.221 | 0.071 | 0.042 | 0.048 | 0.444 | 3.441 | 0.473 | 0.303 | 0.342 |
| 2 | -0.670 | 0.071 | 0.053 | 0.055 | 3.785 | 1.210 | 0.100 | 0.004 | 0.006 |
| 3 | 1.205 | 0.071 | 0.042 | 0.048 | 0.974 | 3.731 | 0.514 | 0.080 | 0.286 |
| 4 | 0.162 | 0.071 | 0.058 | 0.060 | -1.389 | 2.233 | 0.275 | 0.049 | 0.097 |
| 5 | 0.780 | 0.071 | 0.051 | 0.055 | -0.490 | 2.043 | 0.241 | 0.170 | 0.169 |
| 6 | -3.850 | 0.071 | 0.005 | 0.007 | 3.008 | 1.789 | 0.196 | 0.004 | 0.009 |
| 7 | 3.330 | 0.071 | 0.008 | 0.012 | -0.623 | 2.680 | 0.353 | 0.193 | 0.220 |
| 8 | -0.197 | 0.071 | 0.058 | 0.060 | 2.971 | 1.960 | 0.226 | 0.003 | 0.008 |
| 9 | 2.373 | 0.071 | 0.019 | 0.025 | 1.244 | 2.602 | 0.340 | 0.058 | 0.149 |
| 10 | 2.037 | 0.071 | 0.025 | 0.031 | -0.711 | 2.517 | 0.325 | 0.163 | 0.194 |
| 11 | 0.423 | 0.071 | 0.056 | 0.059 | -1.353 | 2.350 | 0.296 | 0.050 | 0.106 |
| 12 | -1.702 | 0.071 | 0.032 | 0.036 | -2.184 | 0.590 | 0.026 | 0.015 | 0.016 |
| 13 | 0.948 | 0.071 | 0.048 | 0.053 | -2.332 | 2.241 | 0.276 | 0.007 | 0.025 |
| 14 | 2.979 | 0.071 | 0.011 | 0.016 | -0.061 | 2.757 | 0.366 | 0.320 | 0.277 |
| 15 | 3.873 | 0.071 | 0.005 | 0.007 | -2.552 | 2.274 | 0.282 | 0.004 | 0.017 |
| 16 | 0.288 | 0.071 | 0.058 | 0.060 | -0.519 | 2.020 | 0.237 | 0.164 | 0.164 |
| 17 | 0.042 | 0.071 | 0.059 | 0.061 | -1.281 | 2.046 | 0.241 | 0.062 | 0.099 |
| 18 | -1.461 | 0.071 | 0.037 | 0.041 | 0.118 | 0.395 | 0.012 | 0.010 | 0.010 |
| 19 | 0.480 | 0.071 | 0.056 | 0.059 | -0.156 | 2.106 | 0.252 | 0.213 | 0.194 |
| 20 | -3.227 | 0.071 | 0.009 | 0.012 | -1.769 | 1.286 | 0.111 | 0.034 | 0.043 |
| 21 | 3.741 | 0.071 | 0.006 | 0.008 | 3.688 | 2.413 | 0.307 | 0.000 | 0.001 |
| 22 | 0.096 | 0.071 | 0.059 | 0.061 | -1.501 | 2.011 | 0.235 | 0.043 | 0.078 |
| 23 | -2.821 | 0.071 | 0.013 | 0.017 | 1.581 | 1.981 | 0.230 | 0.038 | 0.078 |
| 24 | 3.780 | 0.071 | 0.005 | 0.008 | 0.630 | 1.619 | 0.166 | 0.113 | 0.124 |