Accounting for structure in education assessment data using hierarchical models

by

Jillian Dawn Downey

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Ulrike Genschel, Co-major Professor
Mark Kaiser, Co-major Professor
Max Morris
Daniel Nordman
Robert Stephenson

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2017

DEDICATION

To my mother, Denise Lyon, for her never-ending love and encouragement, and for showing me the true meaning of strength.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thank you to all of my friends and family that have helped me through this long journey. A special thank you to my husband, Mitch, for providing me with unconditional support. I also want to thank my committee members, especially Dr. Ulrike Genschel, Dr. Mark Kaiser, and Dr. Dan Nordman for believing in me, even when I didn't always believe in myself.

ABSTRACT

As the field of education continues to grow, new methods and approaches to teaching are being developed with the goal of improving students' understanding of concepts. While research exists showing positive effects for particular teaching methods in small case studies, generalizations to larger populations of students, which are needed to adequately inform policy decisions, can be difficult when using traditional inferential procedures for group comparisons that rely on randomization, replication, and control over relevant factors.

Data collected to compare teaching methods often consists of student level responses, where students are nested within a class, which is typically the experimental unit. Further, for studies in which the scope of inference exceeds individual schools or instructors, we often have classes nested within other factors such as semesters or instructors. In the first part of this dissertation, we explore the consequences of analyzing such data without accounting for the nesting structure. We then show that a hierarchical modeling approach allows us to appropriately account for structure in this type of data. As an illustration, we demonstrate the use of a model-based approach to comparing two teaching methods by fitting a hierarchical model to data from a second course in statistics at Iowa State University.

To fit a hierarchical model to a dataset, the nesting structure must be chosen, a priori. However, with data from an educational setting, there can be instances when the nesting structure is ambiguous. For example, should semesters be nested within instructors or vice versa? In part two of this dissertation, we develop a data-driven diagnostic using moment-based variance estimators to aid in the choice of nesting structure prior to fitting a hierarchical model. We conduct a simulation study to demonstrate the diagnostic's effectiveness and then apply the diagnostic to data from a nationally recognized standardized exam measuring statistical understanding after a first course in statistics. The results from the diagnostic and the subsequent fitted hierarchical model demonstrate

the presence of a difference between the upper level grouping variable that represents the effect of interest. More broadly, this example is intended to highlight the use of hierarchical models for analyzing education data in a way that adequately accounts for variation between students that arises from nested data structures.

CHAPTER 1. INTRODUCTION

There are many new teaching methods and techniques being developed in an attempt to aid student understanding. In introductory college courses, some general examples include flipped classrooms (Winquist and Carlson, 2014), the use of various types of technology, such as clickers (McGowan and Gunderson, 2010), and team-based learning (Clair and Chihara, 2012). An example specific to the discipline of statistics is the use of randomization-based methods to teach introductory statistics courses (Tintle et al., 2011; Maurer and Lock, 2016). Assessing the effectiveness of these new teaching methods and modes of instruction is critical to ensure we are providing students with sufficient presentations of educational material.

Many studies designed to compare modes of teaching or differing pedagogical approaches rely on types of statistical data analysis according to the principles of group, or mean, separation developed for use with designed experiments. While such methods are often straightforward, there are a number of fundamental rudiments on which these methods are based that are difficult to reconcile with the logistical realities faced in designing educational studies. Chief among these are the problems associated with replication and randomization, and the associated scope of inference permitted.

The purpose of this dissertation is to explore the structure in data from an education setting and the use of hierarchical models to account for such structure. Chapter 2 begins by comparing an experimental design approach with a model-based approach to analysis for education data, including a discussion of the difficulties in using the former method in terms of small numbers of experimental units and the permitted scope of inferential statements (Section 2.1). Section 2.2 contains a simulation study demonstrating an increased rate of Type I error when an experimental approach that ignores data structure is used for group comparison analysis of nested data. Section 2.3 introduces hierarchical models as a way to compare groups that takes into account inherent structure in nested data. Finally, in Section 2.4, we demonstrate the presence of structure in education data by fitting a hierarchical model to student-level responses for data from a second course in statistics at Iowa State University.

The focus of Chapter 3 is turned to determining the correct nesting structure for hierarchical models fit to education data. After specifying two different nesting structures and demonstrating the impact each nesting structure has on the marginal variance-covariance matrix in Section 3.1, we calculate moment-based variance estimates for the components of said variance-covariance matrix under each nesting structure in Section 3.2. These moment-based variance estimates are then used in Section 3.3 to develop a data-driven diagnostic for choosing between the two nesting structures. After a short discussion of non-nested models in Section 3.4, the diagnostic is utilized in the analysis of responses from a nationally recognized standardized test which leads to estimation and inference based on a fitted hierarchical model in Section 3.5

CHAPTER 2. STRUCTURE AND VARIATION IN EDUCATION DATA

The comparison of teaching methods has occupied a central place in educational research for many years. In particular, the Scholarship of Teaching and Learning (SOTL) encourages researchers and educators to gather, evaluate and publish evidence about instructional methodology through empirical investigations. Many important motivations exist for an experiment-based approach to education research, such as an educator's intrinsic interest to assess student learning, or a goal to synthesize potential advancements in technology, or even a basic need to adapt to general changes in the landscape of higher education (i.e., higher student enrollments). Due to their comparative nature, these investigations often involve statistical analyses to quantify and contrast the efficacy of different modes of instruction. In the literature, examinations or commentaries have occasionally appeared regarding the statistical practices used in education research (Shaver and Norton, 1980; Geis, 1984; Horton et al., 1993; Walczak et al., 2010). Generally, such critiques have not been complimentary of procedures used by researchers to compare teaching methods, often finding faults in the descriptions of populations and samples, or in random selection and assignment, or in replication and issues of unsubstantiated generalization. There have also been some more recent reviews of educational research methodology, tending to focus on narrow categories of teaching methods and subject disciplines (Waltz et al., 2014; Betihavas et al., 2016; Hainey et al., 2016). These articles often end with a call for additional or better investigations that involve more rigorous statistical methods than used in those studies examined for review.

During the past decade, a number of papers have appeared that either mention or make use of hierarchical models in the analysis of educational data (McGowan and Gunderson, 2010; McGowan, 2011; Niehaus et al., 2014; Chance et al., 2016). Rather than strictly applying concepts from an experimental approach to analysis and inference, one impetus for using statistical models has been a desire to deal with dependence among student-level responses within classes (McGowan, 2011;

Chance et al., 2016). However, there exist broader relations of analysis via hierarchical models to analysis under an experimental approach, which are helpful to understand in examining educational data. These relations have not been fully explored in the context of studies for the comparison of teaching methods, and are also connected to issues regarding the number of classes included and the appropriate scope of inference.

This chapter examines two major difficulties in applying concepts from classical experimental design to the comparison of teaching methods, namely, the appropriate identification of experimental units and the appropriate identification of the scope of inference possible for a given study. In Section 2.1, we discuss these two problems, and contrast the effects they have under experimental and model-based approaches to analysis and inference. Additionally, we address how studies with differing designs are impacted by the above issues to a varying degree depending on sample sizes. We will assert that a fundamental deficiency in the assessment of education data is failing to account for sources of structure in data, rather than using study designs that violate specific assumptions embodied in statistical models. A simulation study in Section 2.2 is provided to illustrate the deleterious effects of ignoring data structure in settings with multiple classes receiving different teaching methods. In Section 2.3 we suggest the use of hierarchical models as a possible vehicle to take proper account of data structure and to avoid the concerns identified in interpreting treatment effects. Section 2.4 further demonstrates the application of a hierarchical model with real education data collected for illustration at our institution, and Section 2.5 offers a discussion and concluding remarks.

2.1 Study Design, Data Structures, and Experimental Concepts

Statistical methods often applied in the comparison of teaching techniques are associated with classical experimental designs. It is common in such comparisons for the concept of experimental units to be misconstrued, with ramifications for what scope of inference is possible. Consider a study in which two teaching methods are used, one in each of two classes. Ignoring any other potential concerns, such as those associated with randomization, this can then be viewed as an unreplicated experiment (Perrett, 2012), because whole classes, not students, directly receive physically independent applications of the treatments and thereby qualify as experimental units (Peck et al., 2012). Following proper experimental reasoning, the study would result in two data values, one for each class (experimental unit). The responses for experimental units may be averages of student-level responses, but students are sampling units, not experimental units, and the comparison of teaching methods boils down to the comparison of the two numbers, one for each class. No analysis using statistical inference is possible or appropriate.

In the example of the previous paragraph in which two classes receive a different (unreplicated) teaching method, one might question if it is possible to take student-level responses to be independent, and conduct a group comparison with a traditional statistical model that leads to a two-sample t-test. This is indeed possible and is likely what most statisticians would recommend in such a situation. For a given teaching method, an assumption that student responses are independent may be motivated as reasonable by using a typical model (e.g., Normal distribution) formulated to compare means of the two groups (classes). We may now employ a statistical comparison, though we are basing that comparison on a statistical model rather than strict experimental concepts. What is most important, however, is that the appropriate scope of inference renders the conclusion applicable to only the two classes examined, which is still anecdotal evidence. Many nearly identical studies with nearly identical results are then needed to inform policy or to suggest general adoption of a given teaching approach. Indeed, a major criticism of Geis (1984) was the failure to define methods completely and precisely, thus prohibiting replication in other studies, which he characterized as "a primary criterion of research."

Now consider another comparative study in which two teaching methods are each applied to a small number of classes. Here again, the proper identification of experimental units is classes, not students. From an experimental viewpoint, the study is no longer unreplicated, though the number of observations for each teaching method (the number of classes receiving said method) might be quite small. A statistical analysis relying on experimental concepts would be possible, but would likely be of low power to detect a difference in teaching methods. One might try to extend the logic from the previous example and consider student-level responses as independent within and among classes, and then make use of a statistical model that ignores the class groupings and perform a two-sample t-test. In this situation, most statisticians would probably conclude that such an analysis is not reasonable, because there may be effects due to class identities that have not been accounted for in the model. That is, even within the same teaching method, it is likely that responses of students in the same class will be more similar than responses of students from different classes. For a defensible analysis to be conducted, this additional structure in the data (differences among classes within teaching methods) should be incorporated into a statistical model. This is often done using a fixed effect for class. The corresponding scope of inference remains the collection of classes actually observed, though that collection is now somewhat larger than the previous situation that included only two classes.

Finally, consider a study in which a reasonably large number of classes receive each of the same teaching method. As the number of classes per teaching method grows, estimation of individual class fixed effects can become cumbersome. Further, other potentially attributing factors are likely to become important variables (e.g., instructors, instructor experience, class time, etc.). In this situation, many statisticians would recommend a model with random class effects. Structure in the data that results from similarity of student responses within classes would now be represented as a variance among classes (i.e., an effect due to variability), rather than a set of small individual differences (i.e., a mean effect). The consequence of random class effects is that student responses in the same class would no longer be independent in the marginal distribution of responses. Why, in our previous examples, was it reasonable to assume that student responses within classes were independent, but now our model produces dependence among those same responses? It would seem that the responses of students within classes should be either dependent or independent, not something that changes being contingent on the number of classes receiving each teaching method. The key to understanding this seeming paradox is that the responses of students within classes are neither independent nor dependent. That is, statistical dependence is not the same as physical dependence. Rather, it is a mathematical property that may or may not be useful in representing structure in sets of data. In all three examples, it is likely that responses of students within classes will be more alike than responses of students from different classes.

In the case of only one class per method, a difference between classes is completely confounded with any difference between teaching methods, and thus the structure of differences between classes and methods coincide. In the case of a small number of classes per method, data structure resulting from differences between classes may be accounted for by a set of constants, numbers that represent the effect of each individually identifiable class. In the case of a greater number of classes per method, however, we lose interest in the identities of individual classes and the data structure induced by differences among classes within method is now accounted for through a combination of random class effects and conditional independence among student responses given those effects. This, in turn, produces dependence in the marginal distribution of all responses. The scope of inference in this situation remains the total set of classes included in the analysis (which is larger than previously considered), unless those classes were randomly selected from some larger population. If this is the case, the scope of inference may be extended beyond the set of observed classes only. At some point, the number of classes included in a modeling analysis plays the same role as the number of replicates involved under an experimental analysis.

To summarize, under an experimental approach assuming independent responses is always appropriate, because responses are defined only for experimental units (rather than sampling units) and the definition of experimental unit hinges on physically independent treatment application. The other critical needs are for complete control over all factors relevant to response values, as well as randomized treatment assignment. Together, these aspects of experimentation guarantee that the only possible source of structure in data is treatment effects. But it is notoriously difficult to faithfully adhere to these experimental requirements in education studies. Modeling unreplicated experiments or small studies using student-level responses can be effective, but reduces the scope of inference. Care must be taken to explicitly account for sources of structure in the data used in order to motivate assumptions of independence among student-level responses. In contrast, modeling larger studies can result in an increased scope of inference, but with data structure now represented through statistical dependencies that result from the use of random effects. In these latter situations, potential causes of data structure (that will need to be accounted for in analysis) can often be identified through nesting of groups. For example, we might have multiple instructors using different teaching methods such that each instructor has responsibility for several classes. In this case, we have a natural nesting of students within classes, and classes within instructors. This implies there will be two components of data structure beyond any possible difference in teaching methods, namely, similarity of students with the same instructor but in different classes, and similarity of students in the same classes.

2.2 Consequences of Ignoring Data Strcture

2.2.1 Variance Estimation in Education Data

In this section, we will illustrate the detrimental effect of ignoring data structure induced by having multiple classes, each taught by a number of instructors for two teaching methods we wish be compare. To define the setting, let Y_{ijkl} be the response for student l ($l = 1, 2, \dots L$), in class k ($k = 1, 2, \dots K$), with instructor j, ($j = 1, 2, \dots J$), using teaching method i ($i = 1, 2$). For simplicity, we assume a completely balanced design in which each teaching method has the same number of instructors (J), each instructor teaches the same number of classes (K), and each class has the same number of students (L). We define the covariance components associated with the marginal distribution of these random variables as follows:

$$\begin{aligned} \text{Var}(Y_{ijkl}) &= a, \\ \text{Cov}(Y_{ijkl}, Y_{ijkl'}) &= b, \text{ for } l \neq l', \\ \text{Cov}(Y_{ijkl}, Y_{ijk'l'}) &= c, \text{ for } k \neq k', l \neq l', \\ \text{Cov}(Y_{ijkl}, Y_{ij'kl'}) &= 0, \text{ for } j \neq j', l \neq l'. \end{aligned} \tag{2.1}$$

In addition, we assume that the response variable at the student level conditioned on the class mean is Normally distributed, that class means conditioned on instructor means are Normally distributed, and that instructor means conditioned on teaching method are Normally distributed.

Typical group comparison techniques consider the difference in means of the response variable under each teaching method (i.e., $\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}$ where $\overline{Y}_{i\ldots}$ denotes the average response associated with the i^{th} teaching method). Statistical inference (tests of hypotheses or confidence intervals) then requires an estimate of the standard error of this difference in sample means, to assess the relative size of the average difference $\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)$. Calculating the standard error without accounting for data structure due to nesting can lead to underestimation of the standard error for the difference in group means. That is, by treating responses as mutually independent and ignoring the natural nesting of students in classes and classes taught by the same instructor, the estimated standard error is smaller than the true value which leads to overstating the significance of any difference in means. This aspect of underestimation is easily and rigorously seen by comparing formulas for standard errors when including or ignoring the data nesting structure (Equations (2.3) and (2.4) to follow).

The general structure of the standard error for the difference in group means, assuming observations between the two groups are independent, is given in Equation (2.2).

$$\begin{aligned}
\sqrt{\mathrm{Var}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} &= \sqrt{\mathrm{Var}\left(\overline{Y}_{1\ldots}\right) + \mathrm{Var}\left(\overline{Y}_{2\ldots}\right) - 2\mathrm{Cov}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} \\
&= \sqrt{\mathrm{Var}\left(\overline{Y}_{1\ldots}\right) + \mathrm{Var}\left(\overline{Y}_{2\ldots}\right)} \\
&= \sqrt{\mathrm{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{1jkl}\right) + \mathrm{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{2jkl}\right)}.
\end{aligned} \tag{2.2}$$

If we ignore data structure and assume that observations within each classroom are independent, as is done in a two-sample t-test or a fixed effects ANOVA, the standard error in Equation (2.2) simplifies as follows,

$$\begin{aligned}
\sqrt{\mathrm{Var}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} &= \sqrt{\mathrm{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{1jkl}\right) + \mathrm{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{2jkl}\right)} \\
&= \sqrt{\frac{JKL}{(JKL)^2}\mathrm{Var}\left(Y_{1jkl}\right) + \frac{JKL}{(JKL)^2}\mathrm{Var}\left(Y_{2jkl}\right)} \\
&= \sqrt{\frac{a}{JKL} + \frac{a}{JKL}} \\
&= \sqrt{\frac{2a}{JKL}},
\end{aligned} \tag{2.3}$$

where a is the variance of a single response as defined in Equation (2.1). However, if we account for the nesting structure outlined in Equation (2.1), the standard error for the difference in group means increases to,

$$\begin{aligned}
\sqrt{\operatorname{Var}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} &= \sqrt{\operatorname{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{1jkl}\right) + \operatorname{Var}\left(\frac{1}{JKL}\sum_{j,k,l} Y_{2jkl}\right)} \\
&= \sqrt{2\left(\frac{a}{JKL} + \frac{(b)(L-1)}{JKL} + \frac{(c)(L)(K-1)}{JKL}\right)} \\
&= \sqrt{\frac{2a}{JKL} + \frac{(2b)(L-1)}{JKL} + \frac{(2c)(L)(K-1)}{JKL}} \\
&> \sqrt{\frac{2a}{JKL}},
\end{aligned} \tag{2.4}$$

where a, b, and c are as defined in Equation (2.1).

A direct comparison of the estimators in Equations (2.3) and (2.4) shows that when we correctly account for the nesting structure reflecting dependence between students, the standard error increases. The implication for ignoring data structure is then quite negative for evaluating potential differences in teaching methods. Not only does an incorrect standard error result, but the dangers for decision making are further compounded by the fact that the standard error when ignoring data structure is smaller than it should be. This increases the chance to falsely conclude there exists a statistically significant difference in teaching method averages $\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)$ compared to when using the correct, larger standard error that accounts for the data structure. In other words, ignoring data structure and underestimating this standard error serves to inaccurately expand the size of the critical region which determines when we can reject the null hypothesis of no treatment differences.

We next present a simulation study to illustrate the effect of disregarding the nesting structure and thus underestimating the standard errors for the difference in group means.

2.2.2 Simulation Study

All analyses in this section were done using R 3.3.2 (R Core Team, 2016). For the purpose of this simulation study, we assume the nesting structure as outlined in Section 2.2.1. Specifically, we consider two teaching methods for comparison, and nested within each teaching method are two instructors, resulting in four instructors in total. Each instructor is assumed to teach four classes, consisting of thirty students each. Thus, this scenario produces data for sixteen classes with a total of 480 students, where we assume a balanced design for simplicity.

Let Y_{ijkl} be the response for student l ($l = 1, 2, \ldots 30$), in class k ($k = 1, 2, 3, 4$), with instructor j, ($j = 1, 2$), using teaching method i ($i = 1, 2$). A hierarchical structure with all Normal distributions was used to simulate the data and is given by Equation (2.5).

$$\begin{aligned} Y_{ijkl}|\beta_{0(ijk)} &\sim \mathrm{N}(\beta_{0(ijk)}, \delta^2 = 16^2), \\ \beta_{0(ijk)}|\mu_{ij} &\sim \mathrm{N}(\mu_{ij}, \tau^2 = 7^2), \\ \mu_{ij}|\lambda_i &\sim \mathrm{N}(\lambda_i, \psi^2 = 4^2), \\ \lambda_1 &= 65, \\ \lambda_2 &= 65. \end{aligned} \tag{2.5}$$

The parameter values chosen for the simulation study are based on an analysis of real data collected from an introductory statistics course taught at the authors' institution. Values for λ_1 and λ_2, representing the mean of the responses under each teaching method, were chosen to reflect no difference between the two methods. Here the variance among students (δ^2) is substantially larger than the variance between class means (τ^2), which in turn is larger than the variance of instructor means (ψ^2). Experience with multiple sets of data suggest the magnitudes of these variances are within the ranges that might be expected in education data. In analogy to a realistic scenario, we might consider the simulated responses, Y_{ijkl}, as representing exam scores, which explains the scale of the responses in this simulation study. (For this reason, as well, values of student responses were technically simulated from a truncated Normal distribution so that all values are between zero and 100, with simulated datasets containing two or fewer truncated values, on average.) We simulated $2,500$ datasets according to the model given in Equation (2.5).

Of primary interest is the Type I error rate (i.e., the proportion of data simulations in which a significant difference is falsely concluded between the two teaching approaches) as a result of ignoring the data structure from the nested responses. For each simulated dataset, we performed a two-sample t-test for assessing the difference between the teaching methods. By assumption, t-tests presume independence between observations from the same group, hence, the estimated standard errors associated with each test should be of the form given in Equation (2.3). Additionally, to consider a model-based approach that accommodates for differences among classes through mean effects (which still misses structure in data related to variability in responses), we also conducted a two-factor analysis of variance assuming a fixed effect for both teaching method and instructor within a given teaching method. As with t-tests, this analysis of variance is based on the assumption that observations within the same group are independent from one another. By applying both testing procedures to all simulated data sets and recording if teaching methods were (erroneously) declared as statistically different at the 0.05 level of significance, we obtained the observed Type I error rates for each testing procedure as reported in Table 2.1.

Table 2.1 Observed proportion of falsely detected significant differences between treatments, i.e., teaching approaches using $\alpha = 0.05$.

	Observed Type I Error Rate	**Nominal Type I Error Rate**
Two sample t-test	0.40	0.05
ANOVA	0.41	0.05

Table 2.1 shows that a failure to account for data structure results in poor Type I error rates. While it is not surprising that neglecting the true data structure may lead to some distortions in analysis and judging significance, the level of distortion is extreme. The observed Type I error rate is 0.40 for the two-sample t-test and 0.41 for the two-factor ANOVA, corresponding to about eight times the size of the intended 0.05 nominal rate. Note that, compared to the two-sample t-test, the more sophisticated two-factor ANOVA does not offer any improvements to the Type I error rate here. Even though this latter testing approach aims to explain data through using fixed-effect terms beyond treatments (i.e., terms for instructor), such terms do not adequately capture the structure

that influences variation in the data. Again, both methods presume independent observations and, when instead student responses within the same treatment group, instructor, or classroom share sources of variability (i.e., are positively correlated), underestimation occurs for the standard error associated with the difference in treatment means. This leads to a larger rejection region and, hence, a higher rejection rate, i.e., we falsely detect a significant difference between treatments more often than should be expected.

For further illustration, we may numerically compare the actual standard errors in this setting under Equations (2.3) and (2.4) (i.e., ignoring the data structure or not). Under the assumption of complete independence among responses, and using the parameter values given in (2.5), the standard error of the estimated difference between the group means is obtained from (2.3) resulting in a value of 1.636 (calculations shown in (2.6)).

$$\begin{aligned}
\sqrt{\text{Var}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} &= \sqrt{\tfrac{2}{JKL}\,(a)} \\
&= \sqrt{\tfrac{2}{240}\,(16^2 + 7^2 + 4^2)} \\
&= \sqrt{\tfrac{321}{120}} \\
&\approx 1.636.
\end{aligned} \tag{2.6}$$

However, as seen in Equation (2.7), when accounting for the variability in responses from the data nesting structure and correctly computing the standard error according to Equation (2.4), we see an increase in its value to 5.512.

$$\begin{aligned}
\sqrt{\text{Var}\left(\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}\right)} &= \sqrt{\tfrac{2}{JKL}\,[a + b(L-1) + c(L)(K-1)]} \\
&= \sqrt{\tfrac{2}{JKL}\,[(\delta^2 + \tau^2 + \psi^2) + (\tau^2 + \psi^2)(L-1) + (\psi^2)(L)(K-1)]} \\
&= \sqrt{\tfrac{2}{240}\,[(16^2 + 7^2 + 4^2) + (7^2 + 4^2)(30-1) + (4^2)(30)(4-1)]} \\
&= \sqrt{\tfrac{3646}{120}} \\
&\approx 5.512.
\end{aligned} \tag{2.7}$$

For the $2,500$ simulated data sets, a histogram of the resulting estimated standard errors for the difference in sample averages when based on Equation (2.3), (i.e., presuming independence among all responses) is shown in Figure 2.1. From the histogram, it is apparent that these estimated standard errors are around the 1.636 value from Equation (2.6) under independence, and at most half the size of the true value of the standard error, 5.512, given in Equation (2.7).

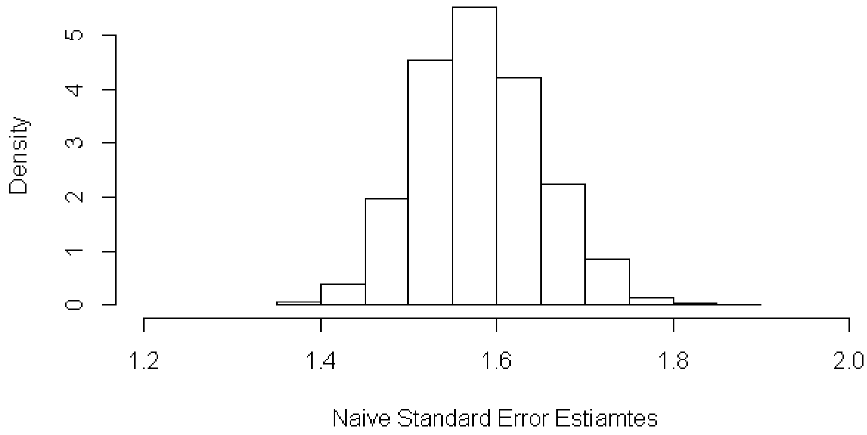


Figure 2.1 Histogram of naive standard error estimates corresponding to the difference in treatment means for all $2,500$ simulated data sets.

In practice, for education data containing similar structure and for which complete independence assumptions may be analogously inappropriate, the determination of standard errors for assessing differences in treatment averages can require estimation of sources of variability, such as the covariance components, $a = \text{Var}(Y_{ijkl})$, $b = \text{Cov}(Y_{ijkl}, Y_{ijkl'})$ and $c = \text{Cov}(Y_{ijkl}, Y_{ijk'l'})$, appearing in Equation (2.7). This is often difficult, if not impossible, to address without some modeling prescriptions. Therefore, we propose using hierarchical models to help explain sources of variability due to data structure and to estimate these covariance-related quantities. Section 2.3 provides a brief discussion of the structure and variation commonly found in data collected in education settings for the purpose of motivating the use of hierarchical models.

2.3 An Alternative Model-Based Approach - Hierarchical Linear Models

2.3.1 Background on Structure and Variation in Data

Data that do not constitute pure noise contain structure, which may be identified as patterns of variation in observations. The classical design-of-experiment approach strives to remove potential sources of structure other than those due to the treatment of interest. In education studies, the difficulty in realizing this ideal has been the thesis of the multiple commentaries, as described in the introduction for this chapter. In the following, we pursue an alternative model-based approach for the analysis of data encountered in education research. Rather than ignoring sources of structure or assuming that these have been adequately removed through design, the model-based approach we present uses hierarchical models to identify and account for structure that may exist in data observed in education studies. The intent is for the analysis and comparison of teaching methods to be profitably conducted based on student-level responses. This approach can have practical advantages when there are enough data to inform some structure, but observations at the classroom level might otherwise be impractical to analyze as required by strict principles of experimental design.

In traditional statistical methodology, covariates are often used helping to explain data variation along a given categorical gradient or multiple gradients. For example, if it is of interest to estimate the effect of semesters in a small study with limited inferential scope, one might explicitly identify these semesters (e.g., semester one or semester two) and treat both as identified covariates with a fixed effect in a statistical model. Response variables measured on classes in the same semester would then share the effect of that semester. In contrast, for data from several semesters where the individual identity of a semester is no longer relevant to the study, a more appropriate choice in modeling the influence of semesters is that of a random effect. Through random effects, the structure in data from multiple semesters is associated with a source of variability, where there is a shared effect of a given semester among classes in that semester. This shared effect manifests as non-zero covariance in a statistical model. Here, the effects of semesters are not represented as a small set of fixed numbers, but rather as a broader range of possibilities for variation. The effect of any single semester is considered a random draw from this broad range of possibilities for variation, represented by an entire distribution. This latter approach to representing various structures that occur in data is that of hierarchical modeling.

If the goal of researchers is to model data at the student level, it is critical that important sources of structure in the data are properly accounted for in the statistical model. It also becomes vital to appropriately identify the relevant inferential scope of the analysis conducted. For a study analyzed in this manner to offer any reasonable amount of generality, it must be large, containing multiple classes within any other grouping factor (e.g., teaching methods or semesters).

2.3.2 Motivating Numerical Comparison

As evidenced in the numerical study of Section 2.2, disregarding structure in the variance of data can lead to dangerously inflated Type I error rates due to inaccurately small standard errors for comparing averages among treatments or teaching methods. This was seen to be true for two-sample t-tests as well as for factor-based ANOVA approaches which expand the use of fixed effects terms, but cannot account for structure influencing the variation in the data. However, hierarchical models can account for nesting structures like those presented in our study in a way that more faithfully reflects the actual standard errors associated with method comparisons.

For illustration, returning to the $2,500$ previously simulated datasets, in addition to performing a two-sample t-test and a two-way fixed effects ANOVA with teaching method and instructor as factors, a hierarchical model was also fit to each dataset using a Bayesian approach. For the hierarchical model, the response for each student in a given class was modeled as a Normal distribution with a class specific mean and variance, as shown in Equation (2.8).

$$Y_{ijkl} \sim \mathrm{N}(\beta_{0(ijk)}, \sigma^2_{ijk}). \tag{2.8}$$

The mixing distributions for the hierarchical model, that is, the distributions assigned to prescribe random effect terms in the model, are given in Equation (2.9).

$$\begin{aligned} \beta_{0(ijk)} &\sim \mathrm{N}(\mu_{ij}, \tau^2), \\ \mu_{ij} &\sim \mathrm{N}(\lambda_i, \psi^2), \\ \sigma^2_{ijk} &\sim \mathrm{IG}(\alpha, \gamma), \end{aligned} \tag{2.9}$$

where IG denotes an Inverse Gamma distribution.

Hence, through Equations (2.8) and (2.9), the mixing distributions determine the nesting structure of the model. Specifically, the mean response of a student in a class ($\beta_{0(ijk)}$) is dependent on a mean effect (μ_{ij}) for the class instructor, where the mean effect of instructor is further connected to a mean effect (λ_i) due to the teaching method used by that instructor in the hierarchy. The variance of classes within instructor (τ^2) was taken to be constant over all instructors in Equation(2.9). Similarly, the variance between instructors within a teaching method (ψ^2) was taken to be constant for both teaching methods.

Equation (2.10) gives the prior distributions used to perform the aforementioned Bayesian analysis using Markov Chain Monte Carlo (MCMC) methods. The distributions were chosen based on conjugacy and the parameter values were selected to make the priors diffuse.

$$\begin{aligned} \lambda_i &\sim \mathrm{N}(M = 0, V^2 = 400), \\ \tau &\sim \mathrm{Unif}(0, A = 20), \\ \psi &\sim \mathrm{Unif}(0, D = 20), \\ \alpha &\sim \mathrm{Unif}(0, J = 4), \\ \gamma &\sim \mathrm{Gamma}(G = 0.01, H = 0.01). \end{aligned} \tag{2.10}$$

Upon fitting the hierarchical model (using R 3.3.2) to each simulated dataset in the numerical study, a corresponding 95% Bayesian credible interval for the mean difference between the two teaching methods ($\lambda_1 - \lambda_2$) was also created, and we recorded the proportion of times, across all data simulations, that such intervals failed to contain zero. Recall that there is no true difference among teaching methods in the simulation (i.e., $\lambda_1 = \lambda_1$ holds in the data generation scheme from Equation (2.5)), so that we are determining the proportion of times that a difference among teaching methods could be (incorrectly) declared using 95% intervals from the hierarchical model-based analysis. Out of the $2,500$ simulated datasets, there were zero instances of this type of false conclusion, which contrasts greatly from the results reported in Table 2.1 where two-sample t-tests or ANOVA-based tests had error rates around 40% instead of the 5% nominal level. The explanation of these results is that tests from Table 2.1, as discussed earlier, fail to account for structure in the data attributed to variation, erroneously treating all observations within as independent. However, hierarchical models incorporate a pattern of variation for the data that reflects the manner in which observations were collected (i.e., the nesting structure) and thus account for the naturally occurring covariance between students within the same class, classes taught by the same instructor, and instructors using the same teaching method. Because hierarchical models are able to account for varying levels of dependencies, or types of variation in a hierarchical formulation, they have the potential to be a valuable tool for comparing teaching methods, provided that education data follow a natural nesting structure.

2.4 Detailed Illustration of Structure and Hierarchical Models in the Analysis of Education Data

Through simulation studies, we examined the consequences of disregarding the data structure attributable to nested patterns in variation (Section 2.2) and demonstrated the ability of hierarchical models to account for this nesting structure, when present, and incorporate multiple levels of variation (Section 2.3). Based on real classroom data, we next provide a more detailed illustration of the use of hierarchical models. This aims to demonstrate the types of nesting structures and patterns in variation that can naturally arise with education data at the level of student responses as well as the application of hierarchical models for capturing such structure. The overall intention is to exhibit the ability of hierarchical models to provide a model-based approach for describing student level data in a situation where classroom level observations are sparse limiting the possibility of any analysis based on experimental design principles. In Section 2.4.1, we describe the data set of interest, while Section 2.4.2 specifies a corresponding hierarchical model for the analysis of these data. Model fitting and parameter estimation are provided in Section 2.4.3, and Section 2.4.4 examines the appropriateness of the hierarchical model through model assessments.

2.4.1 Data Description

For this demonstration, we fit a hierarchical model to data from an introductory course at Iowa State University, corresponding to a second semester of Introductory Business Statistics. The data are comprised of two semesters, Spring 2012 and Spring 2013, where in each semester, two lecture times (defined as lecture one and lecture two) were offered. In Spring 2012 each lecture had four lab sections, while in Spring 2013 lecture one had four lab sections and lecture two had five lab sections. The lab sections available to a student within a given semester are conditional on their preferred lecture time. Once a lecture time is determined, a student's chosen lab section is based on the preferred lab time. The enrollment in lab sections varied from nineteen to forty-one students with an average lab size of thirty-three students. In total, there were 571 students across both semesters in 2012 and 2013.

For both semesters and both lectures times, the same course instructor taught, although lab instructors varied within and across semesters. The course instructor presenting the lecture material had taught the class on several occasions prior to the Spring 2012 semester and, according to the instructor, no substantial changes were made to the course design or presentation of material from 2012 to 2013. For each student, first exam and final exam scores were available, where the final exam performance stands as the main outcome of interest. Students within a semester took identical exams, but exams differed across the two semesters. However, the exams were written by the same instructor and we shall assume these were quantitatively comparable.

2.4.2 Hierarchical Model Formulation

Let Y_{ijkl} denote the response variable of interest, where Y_{ijkl} is the final exam score for student l in lab section k, lecture j, and semester i. Hence, the top of this hierarchy structure is represented by the two semesters ($i = 1, 2$). In this illustration, semester acts as the surrogate treatment and there is an expectation of no difference between semesters in average final exam scores. Within each semester, there are two lectures ($j = 1, 2$) and, within each lecture, there are up to five lab sections ($k = 1, 2, 3, 4, 5$). The final exam scores are standardized over all students, regardless of semester, and are modeled using a Normal distribution with a mean that is dependent on the student's lab section and standardized exam one score. We also model variance in responses as the same for students within a lab section, but potentially different among students across labs. The corresponding model for a student response is given in Equation (2.11) where x_{ijkl} represents a student's (standardized) exam one score.

$$Y_{ijkl} \sim \mathrm{N}(\beta_{0(ijk)} + \beta_1 x_{ijkl}, \sigma^2_{ijk}). \tag{2.11}$$

Note that the model specification in Equation (2.11) is motivated by the nesting structure in the actual data. Histograms of the final exam scores for individual lab sections were examined along with summary statistics. These histograms indicated that the choice of a Normal distribution model of the responses was reasonable. Further, the summary statistics supported the modeling notion of allowing the means and variances to differ across lab sections. As the exam one scores are available, this variable is used as a covariate in the mean structure of final exam score to act as a crude surrogate for student ability and diligence.

The additional sources of variation as well as semester effects are built into the term $\beta_{0(ijk)}$ of the response model (2.11) using mixing distributions which are the same as the mixing distributions in Equation (2.9). The natural nesting structure in the data is taken into account by having the mean effect of the lab section, $\beta_{0(ijk)}$, be dependent on the mean effect of lecture, μ_{ij}. Similarly, the mean of the lecture effect, μ_{ij}, is impacted by a (fixed) mean effect of the semester, λ_i. Recall that, in this illustration, "semester" serves as the "treatment" and we wish to compare the overall semester effects across the two semesters, thus λ_i represents a target mean of interest. In Equation (2.9) for use in (2.11), the variance between labs within a lecture, τ^2, is assumed constant over all lectures and the variance between lectures within a semester, ψ^2, is modeled as the same value for both semesters.

To perform a Bayesian analysis via MCMC methods, prior distributions are placed on λ_i ($i = 1, 2$), β_1, τ, ψ, α, and γ, akin to the analysis in Section 2.3. The prior distributions, shown in Equation (2.12), are chosen based on general observations of student exam scores by the researchers, previous research on Bayesian hierarchical models, and conditional conjugacy. Parameter values are chosen to make the priors diffuse.

$$\begin{aligned}
\lambda_i &\sim \mathrm{N}(M = 0, V^2 = 400),\\
\beta_1 &\sim \mathrm{N}(0, \sigma^2_{\beta_1} = 100),\\
\tau &\sim \mathrm{Unif}(0, A = 20),\\
\psi &\sim \mathrm{Unif}(0, D = 20),\\
\alpha &\sim \mathrm{Unif}(0, J = 4),\\
\gamma &\sim \mathrm{Gamma}(G = 0.01, H = 0.01).
\end{aligned} \tag{2.12}$$

2.4.3 Fitting the Hierarchical Model: Parameter Estimates and Inference

The hierarchical model in (2.11) with mixing distributions given in (2.9) and priors in (2.12) was estimated using a Metropolis within Gibbs sampler written by the authors in R. The Gibbs sampler was run on the data using three sets of starting values. For each set of starting values, 20,000 observations from the posterior distributions of the parameters were collected after discarding a burn-in of 5,000 observations. Trace plots and auto-regressive plots were examined and the scale reduction factor proposed by Gelman and Rubin (1992) was calculated to provide evidence that the three chains were mixing for each individual parameter. These diagnostics indicated that there were no issues in fitting the model.

Tables 2.2 through 2.6 show five quantiles (2.5%, 25%, 50%, 75%, and 97.5%) for the posterior distribution of each parameter based on the 20,000 collected values from one chain. From these posterior distributions, we can compare and contrast the effects of labs, lectures, and semesters on the target student responses, final exam scores.

Based on 95% credible intervals, there were no significant differences between the parameters representing the surrogate treatments, λ_1 and λ_2 (Table 2.2). The 95% credible intervals for λ_1 and λ_2 are $(-1.42, 1.18)$ and $(-1.21, 1.35)$, respectively, which are quite similar and overlap substantially. Posterior distributions showed considerable variability among lecture means, μ_{ij}, across and within semesters (see Table 2.3). For example, the 95% interval for μ_{11} is $(-0.33, -0.04)$ which is completely disjoint from $(-0.01, 0.22)$, the 95% interval for μ_{22}.

Table 2.2 Posterior quantiles for λ_i based on 20,000 observations after a burn-in of 5,000.

Parameter	2.5%	25%	50%	75%	97.5%
λ_1	-1.42	-0.20	-0.08	0.05	1.18
λ_2	-1.21	-0.04	0.08	0.20	1.35

Table 2.3 Posterior quantiles for μ_{ij} based on 20,000 observations after a burn-in of 5,000.

Parameter	2.5%	25%	50%	75%	97.5%
μ_{11}	-0.33	-0.24	-0.19	-0.14	-0.04
μ_{12}	-0.11	-0.01	0.04	0.09	0.18
μ_{21}	-0.09	0.00	0.05	0.09	0.18
μ_{22}	-0.01	0.07	0.11	0.15	0.22

Summary values of posterior distributions for $\beta_{0(ijk)}$ presented in Table 2.4 are similar within lecture block j for a given semester i. Focusing on lecture 1 in semester 1 we can see that the 95% credible intervals are nearly identical for the four lab sections. Note that the quantiles for $\beta_{0(223)}$ (the mean response for lab section three with lecture two in semester two) do appear somewhat elevated for that particular lecture block, but there remains considerable overlap with average responses for students among other lab sections for this same lecture period in the same semester.

Table 2.4 Posterior quantiles for $\beta_{0(ijk)}$ based on 20,000 observations after a burn-in of 5,000.

Parameter	2.5%	25%	50%	75%	97.5%
$\beta_{0(111)}$	-0.37	-0.25	-0.20	-0.14	-0.04
$\beta_{0(112)}$	-0.33	-0.24	-0.18	-0.13	-0.03
$\beta_{0(113)}$	-0.35	-0.25	-0.20	-0.14	-0.05
$\beta_{0(114)}$	-0.35	-0.25	-0.19	-0.14	-0.04
$\beta_{0(121)}$	-0.10	0.00	0.05	0.10	0.21
$\beta_{0(122)}$	-0.15	-0.03	0.03	0.08	0.19
$\beta_{0(123)}$	-0.11	-0.00	0.05	0.10	0.22
$\beta_{0(124)}$	-0.12	-0.02	0.04	0.09	0.19
$\beta_{0(211)}$	-0.10	0.01	0.06	0.11	0.24
$\beta_{0(212)}$	-0.11	-0.01	0.04	0.09	0.18
$\beta_{0(213)}$	-0.10	0.00	0.05	0.10	0.22
$\beta_{0(214)}$	-0.12	-0.01	0.04	0.08	0.18
$\beta_{0(221)}$	-0.13	0.03	0.09	0.13	0.22
$\beta_{0(222)}$	-0.07	0.05	0.09	0.14	0.22
$\beta_{0(223)}$	0.04	0.11	0.17	0.23	0.36
$\beta_{0(224)}$	-0.04	0.06	0.11	0.15	0.24
$\beta_{0(225)}$	-0.03	0.07	0.11	0.15	0.23

The estimates in Table 2.5 show that there is larger between-lecture variability, represented by ψ, as compared to the between-lab variability, τ. Further, the estimate for the slope parameter, β_1, for the covariate in the model (2.11) is positive and large, indicating that students who performed well on the first exam tended to also perform well on the final. A strong positive relationship between first and final exam score is not unexpected. Based on the fitted model, we see that for every point above the mean on exam one, a student is estimated to be 0.7 points above the mean on the final exam, on average.

Table 2.5 Posterior quantiles for β_1, τ, ψ, α, and γ based on 20,000 observations after a burn-in of 5,000.

Parameter	2.5%	25%	50%	75%	97.5%
β_1	0.63	0.67	0.69	0.71	0.75
τ	0.00	0.03	0.05	0.08	0.16
ψ	0.03	0.13	0.25	0.52	3.91
α	2.37	3.25	3.60	3.82	3.99
γ	1.00	1.42	1.63	1.84	2.24

Finally, Table 2.6 shows the posterior summaries for the within-class variances, σ^2_{ijk}. Taking the posterior medians (i.e., the fiftieth percentiles in Table 2.6) as point estimates, these values indicate some differences among variability within a lab, though these differences do not appear to be systematic and the point estimates of variances tend to share common ranges of possibilities. This aspect supports the modeling of these within-class variances as random variables drawn independently from a common underlying source distribution (given by $\sigma^2_{ijk} \sim \mathrm{IG}(\alpha, \gamma)$ in Equation (2.9)). Comparing the Table 2.5 medians of the 20,000 posterior draws for between-lab ($\tau^2 = 0.0025$) and between-lecture ($\psi^2 = 0.0625$) variances to the estimates for within-class variances in Table 2.6 (values ranging from 0.35 to 0.83), we see that the variability between students within the same lab dominates the other two sources of variability. This agrees with the intuition that natural variability among students might be greater than variability attributable to differing times for lab or lecture periods.

Table 2.6 Posterior quantiles for σ^2_{ijk} based on 20,000 observations after a burn-in of 5,000.

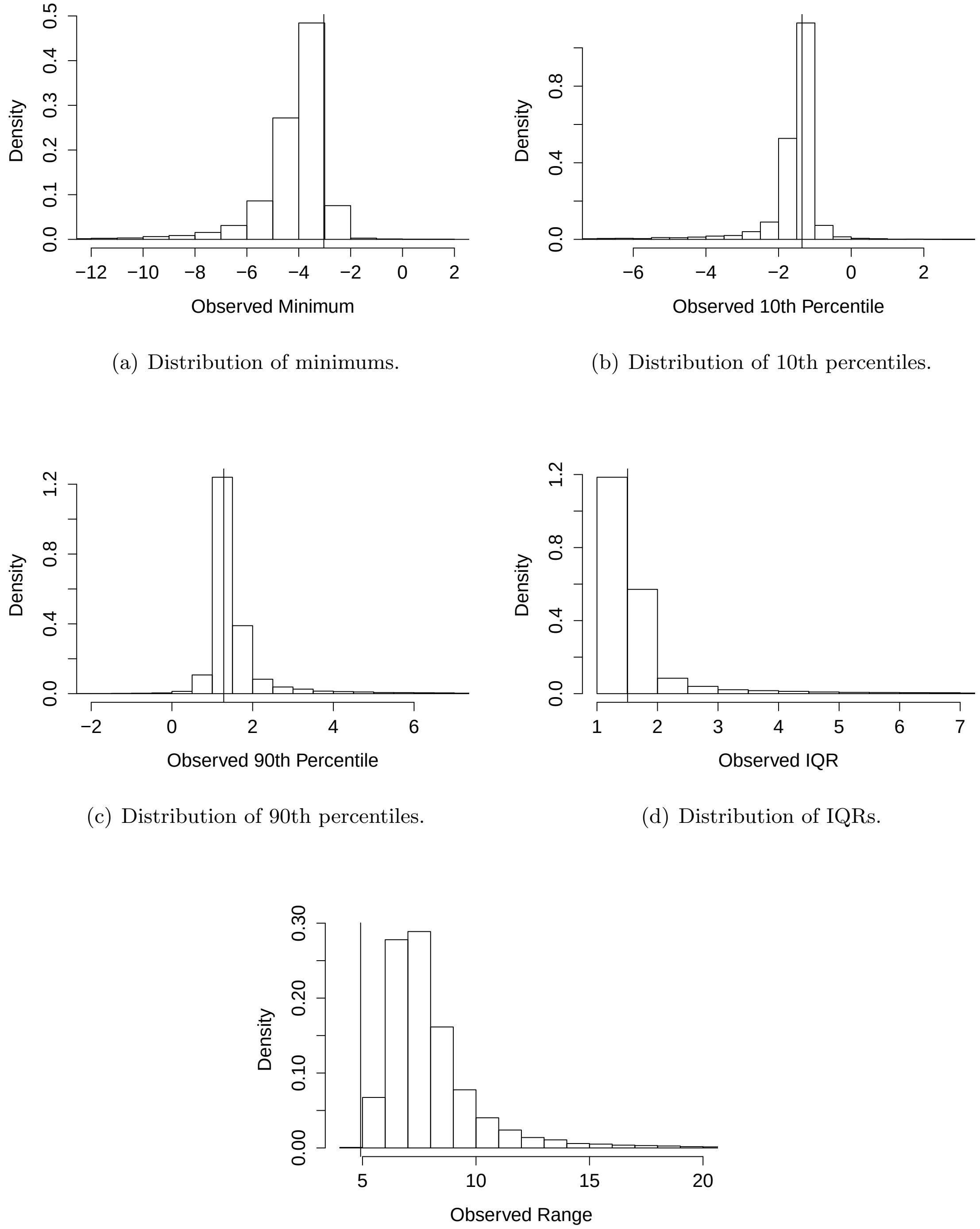
Parameter	2.5%	25%	50%	75%	97.5%
σ^2_{111}	0.44	0.57	0.66	0.78	1.06
σ^2_{112}	0.37	0.47	0.54	0.63	0.86
σ^2_{113}	0.24	0.32	0.38	0.44	0.62
σ^2_{114}	0.36	0.46	0.53	0.61	0.82
σ^2_{121}	0.29	0.37	0.43	0.50	0.69
σ^2_{122}	0.28	0.38	0.45	0.54	0.79
σ^2_{123}	0.42	0.54	0.62	0.72	0.97
σ^2_{124}	0.27	0.36	0.42	0.50	0.69
σ^2_{211}	0.24	0.34	0.41	0.50	0.75
σ^2_{212}	0.27	0.35	0.41	0.47	0.64
σ^2_{213}	0.31	0.42	0.50	0.59	0.86
σ^2_{214}	0.33	0.42	0.49	0.57	0.78
σ^2_{221}	0.54	0.71	0.83	0.96	1.34
σ^2_{222}	0.38	0.48	0.56	0.65	0.88
σ^2_{223}	0.26	0.34	0.39	0.46	0.62
σ^2_{224}	0.32	0.42	0.48	0.56	0.75
σ^2_{225}	0.23	0.30	0.35	0.40	0.55

2.4.4 Model Assessments

In Sections 2.4.2 and 2.4.3 we developed and fit a hierarchical model to account for sources of nested structure in student-level responses concerning final exam scores. We will next address the adequacy of the model by further examining the fitted model to assess whether this model is compatible with the data. The model assessment is based on observations simulated from the posterior predictive distribution of the final exam responses, Y_{ijkl}. For each dataset simulated from the posterior predictive distribution (a description of the procedure is found in Appendix A), five prototypical summaries were calculated; the minimum observation, the tenth percentile, the ninetieth percentile, the inter-quartile range (IQR), and the range. These statistics were compared to the corresponding (true) values obtained from the original data on test scores. Histograms of each statistic calculated from the posterior predictive distributions are shown in Figure 2.2 with the vertical line representing the observed value of a statistic from the original dataset. Note that the range of all of the histograms in Figure 2.2 has been restricted to better visualize the corresponding distributions of these statistics. Full histograms can be found in Appendix B.

The plots corresponding to the minimum, the tenth percentile, the ninetieth percentile and the IQR in Figure 2.2 show that the model is appropriate as a data generating mechanism with respect to these four chosen characteristics, as the vertical line (the observed value of a statistic from the original data) falls within the high density area for each histogram. It appears that the model fails to adequately capture the range of the original dataset. The Normal distribution used in the data model (Equation 2.11) does not reflect that the maximal final exam scores are bounded, suggesting the use of a truncated Normal model in future analyses. The minimal final exam scores are also bounded, however the use of a non-truncated Normal model does not appear to be an issue in the posterior predictive assessment using the smallest data value as a characteristic of interest. This is expected due to the fact that students are more likely to obtain high scores (close to the maximum possible score) than low scores (close to the minimum possible score).

Simulation-based model assessment was also used to verify that the model correctly captures the basic relationships among responses from students within the same lab, lecture, and semester. Again, we simulated 20,000 datasets from the posterior predictive distribution. For each dataset, a simple linear regression model was fit for each lab individually using ordinary least squares, where the simulated values from the posterior predictive distribution served as the response and the observed exam one scores as the covariate, resulting in seventeen slopes and seventeen intercepts. The variance of the slopes and the variance of the intercepts were obtained so that each posterior predictive dataset yielded two variance estimates. For the the 20,000 intercept variance calculations



(a) Distribution of minimums.

(b) Distribution of 10th percentiles.

(c) Distribution of 90th percentiles.

(d) Distribution of IQRs.

(e) Distribution of ranges.

Figure 2.2 Distributions of five different quantities from simulated posterior predictive data sets.

from the posterior predictive distributions, the probability of being more extreme than the observed variance of intercepts was 0.1155. Similarly, for the 20,000 slope variance calculations from the posterior predictive distributions, the probability of being more extreme than the observed variance of slopes was 0.4182. Both probabilities indicate that the model generates data comparable to the original dataset with respect to the variance of the slopes and intercepts for a simple linear regression model fit to individual lab sections.

Finally, we used Kolmogorov-Smirnov statistics based on the empirical distributions of $\widehat{\beta}_{0(ijk)}$ and $\widehat{\beta}_{1(ijk)}$ to assess model fit (procedure outlined in Appendix C). Of the 100 simulated statistics, forty-three were more extreme than the true value for the intercept from the original data. For the slope, thirty of the 100 simulated Kolmogorov-Smirnov statistics were more extreme than the true value from the original data. These probabilities again indicate that the model is adequate as a data-generating mechanism with respect to the distribution of the slopes and intercepts from simple linear regression models fit to individual lab sections.

2.5 Concluding Remarks

Education data contain structure and, as demonstrated in Section 2.2, the consequence of not accounting for such structure includes an inflated Type I error rate leading to an increased number of incorrect inferential statements about the difference between two teaching methods. Hierarchical models are able to account for naturally nested sources of variability and capture the inherent structure in student-level responses. This model-based approach allows for an informative statistical analysis of the data that might otherwise not be feasible due to limited classroom level observations (experimental units). The scope of inference, or the generalization of statistical inference, is always greatest when there are a large number of classrooms. Nevertheless, when the number of classroom-level observations is small, while it may not be possible to improve the scope of inference, it is still possible to conduct an informative model-based evaluation of teaching methods by returning to student-level observations and appropriately accounting for data structure in the model formulation.

Further, hierarchical models can be used to assess educational outcomes while accounting for heterogeneous student populations, and can be applied across multiple classes, instructors, and schools. Analysis of the dataset from an introductory statistics course at Iowa State University performed in Section 2.4 demonstrates that the variance components corresponding to the nested data structure are present in the data, and are appropriately captured by the hierarchical model. In this example, the top level of the hierarchy corresponded to two semesters of instruction for which we, a priori, expected no differences to exist. In applications concerned with evaluating the efficacy of different pedagogical approaches, this level of the hierarchy would correspond to different teaching methods.

In the data from Section 2.4, a clear nesting structure exists. However, not all datasets will have levels so easily nested. For example, if more than one instructor teaches a class over two or more semesters, instructor could be considered nested within semester, or semester nested within instructor. At this point, there is no straightforward methodology available to determine the appropriate nesting structure. The relationship between nesting structures and covariance matrices and the potential use of variance elements in determining an appropriate nesting structure is the subject of ongoing investigation.

CHAPTER 3. SELECTING HIERARCHICAL MODEL STRUCTURES IN THE ANALYSIS OF EDUCATION DATA

Hierarchical models can be used to account for structure in education data that arises from combinations of classes, instructors, semesters, institutions, teaching methods, and other relevant factors. Not accounting for such structure by assuming independence and identical distribution of responses across levels of a hierarchy typically leads to inflated type I error rates when testing, for example, for differences in teaching methods (Section 2.2). When data structure is believed to arise from a nesting of factors, such as classes within institutions, each level of a hierarchical structure can be specified as conditionally independent, given values of random parameters (typically means) that follow distributions over higher levels of the model. Section 2.4, for example, considered a model in which students were nested within labs, which were in turn nested within lecture, which were themselves nested within semester. Given an overall mean for a semester, each lecture then had its own response mean independent of other lectures and, given a lecture mean, each lab had its own response mean independent of other labs. Further, given a lab mean, student responses were independent within that lab section. The use of conditionally independent random variables in such a nested structure is to produce certain dependencies in the marginal distribution of all responses. Thus, for the example just described, student responses are positively correlated among students in the same lab, in the marginal distribution of all responses.

If the nesting structure chosen for a particular model is an adequate representation of the forces acting to produce responses, the dependencies induced in the marginal distribution of responses will capture a large portion of the uncontrolled structure in the data, and inferences about possible differences among upper level factors, such as teaching methods, will be more accurate. In some cases, however, it is not clear as to what an appropriate nesting structure might be. For example, when multiple instructors each teach classes in multiple semesters, would it be appropriate to consider instructors as nested within semesters, or would it be better to consider semesters as nested within instructors? Or, alternatively, would it be best to use a model that contains effects for both instructors and semesters, but in a non-nested manner?

Our objective in this chapter is to develop a data-driven diagnostic that can help guide model choice, specifically identifying the most appropriate nesting structure. It is important to note the context within which our following work has relevance. Hierarchical models are most useful in situations that involve a reasonably large number of groups of responses. In demonstrating the value of our developed diagnostic, we will work with data containing nine, thirty, or 100 groups. We consider nine to be smaller than what we would like for use of a hierarchical model, thirty to be perhaps adequate but still less than ideal, and 100 to be approaching the types of situations for which hierarchical models are well suited. Thus, our work is not intended to address the analysis of educational studies that involve only a few individual groups of students, such as classes. For the majority of this presentation, we will consider responses that are grouped by three factors, with one of those factors containing levels among which we wish to test for differences. That factor will be considered fixed in our models. For each level of this factor we will have two additional factors that will be nested within each model, and will be modeled as arising from distributions for conditionally independent variables. Our diagnostic will concern the nesting structure of these two random factors. All distributions will be specified as Normal, and variances will be assumed constant at each level of the hierarchy.

The remainder of this chapter is organized as follows. In Section 3.1, we specify models for two nesting structures, and present the implications of these models for the full variance-covariance matrix of marginal distributions. In Section 3.2 we consider moment-based estimators of the components of those covariance matrices, which will serve as the basis for developing a data-driven diagnostic to assist in model selection outlined in Section 3.3. Additionally, Section 3.3 presents a simulation study demonstrating the usefulness of the developed diagnostic. Section 3.4 contains a brief consideration of non-nested model structures and how they are related to the proposed diagnostic. An application of the diagnostic is presented in Section 3.5 in the analysis of a large national data set. Section 3.6 concludes this chapter with a discussion and potential future research directions.

3.1 Model Formulations and Covariance Comparisons

Suppose that a response variable of interest is measured at the student level, that is, observational units are represented by individual students in a study for which we wish to compare two teaching methods. We assume students are grouped into classes within certain semesters, and that each class has a single instructor. Thus, in addition to knowing the response of each student, we also know the instructor of each student and in which semester the student received instruction. For simplicity of presentation, we assume that each instructor was responsible for only one class in any given semester, and that one class was taught with a given method for the entire semester. We note that some studies may contain multiple classes taught by the same instructor within a given semester. While our methodology can be applied to these situations as well, we will focus on the simpler case (one class per instructor and semester combination) in the following. As mentioned in the beginning of Chapter 3, the situations we envision here are ones in which a large number of classes are taught by a large number of instructors, such as might occur with a state or nationwide administered measurement instrument, like a standardized exam.

Notation that will be used to formulate models with two different nesting structures is as follows. Let Y_{ijkl} be the response of interest for student l ($l = 1, 2, \ldots, L_{ijk}$) with instructor k ($k = 1, \ldots, K_i$) in semester j ($j = 1, \ldots, J_i$) using teaching method i ($i = 1, 2$). This notation varies slightly from that in Chapter 2 as the number of instructors using teaching method one (denoted K_1) is now allowed to vary from the number of instructors using teaching method two (denoted K_2). Similarly, the number of semesters using teaching method one (denoted J_1) is allowed to vary from the number of semesters using teaching method two (denoted J_2). Finally, the number of students in a class (given a teaching method, instructor, and semester) is denoted as L_{ijk}, which may vary among classes, where the number of classes is equal to the number of unique method, instructor, and semester combinations.

Consider a model in which students are nested within instructors, instructors are nested within semesters, and semesters are nested within teaching methods. Equation (3.1) shows the formulation of such a model, which we will henceforth refer to as Model 1. Under Model 1, students' responses, Y_{ijkl}, are modeled as Normally distributed within a class (instructor, semester, and teaching method combination), and are taken to be conditionally independent. Each class is allowed to have a different mean, θ_{ijk}, but we assume a common variance, δ^2, for all classes. Similarly, the class means, θ_{ijk}, are Normally distributed with an expected value, μ_{ij}, specific to a given semester and teaching method, but with a constant variance, τ^2, and are also assumed to be conditionally independent. Finally, the semester means, μ_{ij}, are Normally distributed with an expected value that depends on the teaching method λ_i, but with the same variability, ψ^2. Using conventional notation, Model 1 may be represented as,

$$\begin{aligned} Y_{ijkl}|\theta_{ijk} &\sim \mathrm{N}(\theta_{ijk}, \delta^2), \\ \theta_{ijk}|\mu_{ij} &\sim \mathrm{N}(\mu_{ij}, \tau^2), \\ \mu_{ij}|\lambda_i &\sim \mathrm{N}(\lambda_i, \psi^2). \end{aligned} \tag{3.1}$$

Next, we consider a model in which the nesting of instructors and semesters is reversed from Model 1, that is, students are nested within semesters, semesters are nested within instructors, and instructors are nested within teaching methods. Model 2 (given by Equation (3.2)) shows the formulation of a model in which semesters are nested within instructors. Note that Model 2 keeps the same indexing as Model 1 to allow for direct comparison. That is, j continues to index semesters, and k continues to index instructors. As for Model 1, in Model 2 student responses, Y_{ijkl}, are Normally distributed within a class (instructor, semester, and teaching method combination) where each class has a different mean, given by θ_{ijk}, but a common variance, δ^2. However, in contrast with Model 1, in Model 2 class means, θ_{ijk}, are Normally distributed with an expected value that varies depending on instructor and teaching method (rather than semester and teaching method). This expectation is denoted by α_{ik} and the variance of the class means, ϵ^2 is assumed to be constant. The instructor means, α_{ik}, are Normally distributed with an expected value that depends on teaching method, λ_i, but with the same variance, σ^2.

$$\begin{aligned} Y_{ijkl}|\theta_{ijk} &\sim \mathrm{N}(\theta_{ijk}, \delta^2), \\ \theta_{ijk}|\alpha_{ik} &\sim \mathrm{N}(\alpha_{ik}, \epsilon^2), \\ \alpha_{ik}|\lambda_i &\sim \mathrm{N}(\lambda_i, \sigma^2). \end{aligned} \tag{3.2}$$

By keeping the indexing the same and calculating elements of the covariance matrix of the marginal distribution of responses, we can highlight the differences in the data structures between Model 1 and Model 2. The non-zero elements of the covariance matrix under Model 1 are given by Expression (3.3) while the non-zero elements of the covariance matrix under Model 2 are given in Expression (3.4). Of particular interest are the values of $\mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'})$ and $\mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'})$. $\mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'})$ represents the covariance between two students taught using the same teaching method in the same semester, but by different instructors. The value of $\mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'})$ is ψ^2 under Model 1 , but zero under Model 2. Similarly, the covariance between two students with the same instructor and teaching method, but different semesters, given by $\mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'})$, is 0 assuming the structure of Model 1, while under Model 2 this value is σ^2.

$$\begin{aligned} \mathrm{Var}(Y_{ijkl}) &= \psi^2 + \tau^2 + \delta^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijkl'}) &= \psi^2 + \tau^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'}) &= \psi^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'}) &= 0. \end{aligned} \tag{3.3}$$

$$\begin{aligned} \mathrm{Var}(Y_{ijkl}) &= \sigma^2 + \epsilon^2 + \delta^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijkl'}) &= \sigma^2 + \epsilon^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'}) &= 0, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'}) &= \sigma^2. \end{aligned} \tag{3.4}$$

To develop a diagnostic useful in model selection, we consider purely data-driven estimation of the elements of these covariance matrices. In particular, we are interested in the comparison of the estimated values of $\mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'})$ and $\mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'})$, as these may indicate which covariance structure is in greater concert with the observed data.

3.2 Estimating Covariances

Typical sample covariances cannot be computed when the available data contain no replication of the overall groupings, that is, when the data represent one realization of a vector-valued random variable. To estimate the covariance elements from a single vector of observations, we make use of the fact that expected values of sample variances of responses within different groupings take the form of linear combinations of elements of the covariances given in (3.3) and (3.4) and obtain moment-based variance estimates. Recall that Y_{ijkl} corresponds to the response of interest for student l with instructor k in semester j using teaching method i. Under this generic setting, specific sample variance calculations of interest are defined as follows.

Let s^2_{ijk} denote the sample variance of all observations with the same method (i), semester (j), and instructor (k), i.e.,

$$s^2_{ijk} = \widehat{\underset{l}{\mathrm{Var}}}(Y_{ijkl}) = \left(\frac{1}{L_{ijk} - 1}\right) \sum_l (Y_{ijkl} - \bar{Y}_{ijk.})^2, \tag{3.5}$$

where L_{ijk} is the number of students in the class with instructor k, in semester j, using teaching method i, Y_{ijkl} represents the response of student l in the class with instructor k, in semester j, using teaching method i, and $\bar{Y}_{ijk.}$ is the average student response for all students with teaching method i in semester j taught by instructor k.

Next, let s^2_{ij} denote the sample variance of all observations within the same same method (i) and semester (j), i.e.,

$$s^2_{ij} = \widehat{\underset{kl}{\mathrm{Var}}}(Y_{ijkl}) = \left(\frac{1}{\sum_k L_{ijk} - 1}\right) \sum_k \sum_l (Y_{ijkl} - \bar{Y}_{ij..})^2, \tag{3.6}$$

where $\bar{Y}_{ij..}$ is the average student response for all students with teaching method i in semester j.

Similarly, let s^2_{ik} represent the sample variance of all observations within the same same method (i) and instructor (k), i.e.,

$$s^2_{ik} = \widehat{\underset{jl}{\mathrm{Var}}}(Y_{ijkl}) = \left(\frac{1}{\sum_j L_{ijk} - 1}\right) \sum_j \sum_l (Y_{ijkl} - \bar{Y}_{i.k.})^2, \tag{3.7}$$

where $\bar{Y}_{i.k.}$ is the average student response for all students with teaching method i and instructor k.

Finally, let s^2_{ijkl} denote the ample variance of all observations, i.e.,

$$s^2_{ijkl} = \underset{ijkl}{\widehat{\text{Var}}}(Y_{ijkl}) = \left(\frac{1}{\sum_i \sum_j \sum_k L_{ijk} - 1}\right) \sum_i \sum_j \sum_k \sum_l (Y_{ijkl} - \bar{Y}_{....})^2, \tag{3.8}$$

where $\bar{Y}_{....}$ is the average student response for all students.

In addition to the sample variance calculations given in (3.5) through (3.8), the values of M_1 and M_2 (see (3.9) and (3.10)) are also needed to obtain the moment-based variance estimates.

We use M_1 to denote the sample average of all observations in teaching method one, i.e.,

$$M_1 = \frac{1}{\sum_j \sum_k L_{1jk}} \left(\sum_j \sum_k \sum_l Y_{1jkl} \right), \tag{3.9}$$

where L_{1jk} is the number of students in the class with instructor k, in semester j, using teaching method one and Y_{1jkl} represents the response of student l in the class with instructor k, in semester j, using teaching method one.

Similarly, let M_2 represent the sample average of all observations in teaching method two, i.e.

$$M_2 = \frac{1}{\sum_j \sum_k L_{2jk}} \left(\sum_j \sum_k \sum_l Y_{2jkl} \right), \tag{3.10}$$

where L_{2jk} is the number of students in the class with instructor k, in semester j, using teaching method two and Y_{2jkl} represents the response of student l in the class with instructor k, in semester j, using teaching method two.

To obtain moment-based variance estimates, we must solve a system of equations in which the theoretical expected values of the sample variances are set equal to the observed sample variances and the theoretical expected values of the sample means are set equal to the observed sample means. The expected values of the sample variances and means depend on the number of students per class as well as the number of instructors and the number of semesters per teaching method. Thus, let us consider the generic situation described at the beginning of this section in which the number of students per class is allowed to vary as are the number of semesters and instructors per teaching method. Under this setting, we have the following system of equations:

$$
\begin{aligned}
\mathrm{E}(M_1) &= \lambda_1, \\
\mathrm{E}(M_2) &= \lambda_2, \\
\mathrm{E}(s^2_{ijk}) &= \delta^2, \\
\mathrm{E}(s^2_{ij}) &= \delta^2 + a(\tau^2), \\
\mathrm{E}(s^2_{ik}) &= \delta^2 + b(\epsilon^2), \\
\mathrm{E}(s^2_{ijkl}) &= \left(\tfrac{1}{d}\right)\left[f(\lambda_1^2) + g(\lambda_2^2) + h(\lambda_1\lambda_2) + n(\delta^2) + p(\tau^2) + q(\psi^2)\right] \\
&= \left(\tfrac{1}{d}\right)\left[f(\lambda_1^2) + g(\lambda_2^2) + h(\lambda_1\lambda_2) + n(\delta^2) + p(\epsilon^2) + r(\sigma^2)\right].
\end{aligned}
\tag{3.11}
$$

The values of $a, b, d, f, g, h, n, p, q$ and r in (3.11) represent coefficients that depend on the number of students, instructors, and semesters in a particular dataset. The preceding system of equations can be solved in an iterative manner to obtain unbiased estimates for the elements of the variance-covariance matrices based on Model 1 (given in (3.3)) and Model 2 (given in (3.4)).

Consider the situation in which $l = 1, 2, \ldots, L_{ijk}$, $k = 1, \ldots, K_i$, and $j = 1, \ldots, J_i$. Here, K_1 represents the number of instructors in teaching method one and K_2 represents the number of instructors in teaching method two. Similarly, J_1 and J_2 represent the number of semesters in teaching method one and two, respectively. Define C to be the total number of distinct classes, where each class has a unique method, semester, and instructor combination. Under this generic setting, the moment-based estimators for the two models are given in Equation (3.12). The specific forms of the coefficients ($a, b, d, f, g, h, n, p, q$ and r) for the general setting and for a completely balanced design are given in Appendix D and Appendix E, respectively. Note that this estimating technique allows for negative values of variance components (most likely to occur at the data level), which proves to be an informative characteristic of these moment-based estimators for use as a model selection diagnostic.

$$\begin{aligned}
\widehat{\lambda}_1 &= M_1, \\
\widehat{\lambda}_2 &= M_2, \\
\widehat{\delta}^2 &= \left(\tfrac{1}{C}\right) \textstyle\sum_i \sum_j \sum_k \left(s^2_{ijk}\right), \\
\widehat{\tau}^2 &= \left(\tfrac{1}{J_1+J_2}\right) \textstyle\sum_i \sum_j \left[\tfrac{1}{a}\left(s^2_{ij} - \widehat{\delta}^2\right)\right], \\
\widehat{\epsilon}^2 &= \left(\tfrac{1}{K_1+K_2}\right) \textstyle\sum_i \sum_k \left[\tfrac{1}{b}\left(s^2_{ik} - \widehat{\delta}^2\right)\right], \\
\widehat{\psi}^2 &= \left(\tfrac{1}{q}\right) \left[d(s^2_{ijkl}) - f(\widehat{\lambda}^2_1) - g(\widehat{\lambda}^2_2) - h(\widehat{\lambda}_1\widehat{\lambda}_2) - n(\widehat{\delta}^2) - p(\widehat{\tau}^2)\right], \\
\widehat{\sigma}^2 &= \left(\tfrac{1}{r}\right) \left[d(s^2_{ijkl}) - f(\widehat{\lambda}^2_1) - g(\widehat{\lambda}^2_2) - h(\widehat{\lambda}_1\widehat{\lambda}_2) - n(\widehat{\delta}^2) - p(\widehat{\epsilon}^2)\right].
\end{aligned} \tag{3.12}$$

3.3 A Diagnostic of Nesting Structure

Due to the iterative nature of the moment-based estimators developed in Section 3.2, it is possible to obtain negative variance estimates. We conjecture that negative moment-based variance estimates are more likely to occur under the incorrect nesting structure due to the incorrect grouping of less homogeneous students. Thus, we propose to use this property as a model-choice diagnostic.

3.3.1 Nested Hierarchies and Homogeneous Data Groupings

As an illustration, consider a dataset that follows the structure of Model 1, which assumes that instructors are nested within semesters. For simplicity, we assume we have three semesters and two instructors per semester for each of two teaching methods. Further, we assume that any given teaching method, semester, and instructor combination only has one class for a total of twelve classes, six per teaching method. A picture of this nesting structure for a given teaching method (say Method 1) is shown in Figure 3.1.

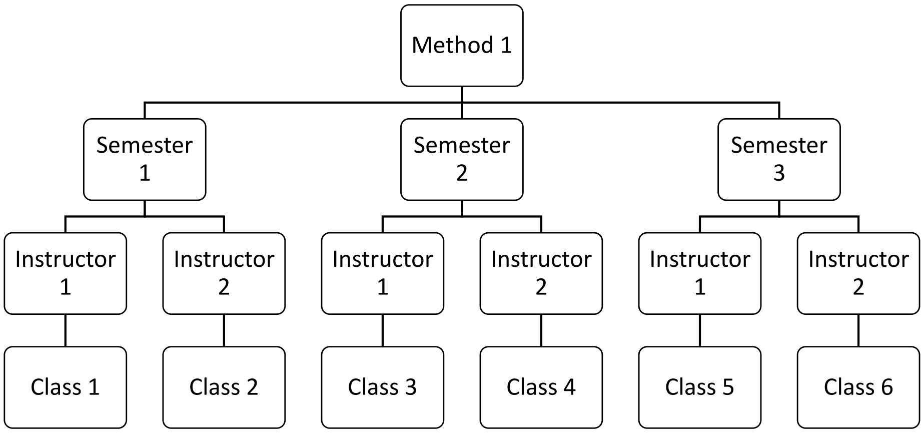


Figure 3.1 The nesting structure of Model 1 within a given teaching method for a simple case.

The moment-based variance estimators utilize different groupings of students to try to find the appropriate hierarchical structure in a data-driven manner. If we assume the correct model structure (Model 1) for our illustration, we estimate the values of δ^2, τ^2, and ψ^2 using Equation (3.12). Of specific interest is the calculation of $\hat{\tau}^2$ which is computed as a linear combination of $\hat{\delta}^2$ and all s_{ij}^2's where s_{ij}^2 (given by Equation (3.6)) represents the variance of observations within teaching method i in semester j. Continuing to focus on one teaching method in our illustration, a picture of the groupings used to calculate the s_{ij}^2's (and thus $\hat{\tau}^2$) is shown in Figure 3.2. This picture shows that $\hat{\tau}^2$ is found by taking the sample variance of all students with the same class color, then calculating a linear combination of these variances over the three different color groups.

In contrast, if we assume the incorrect nesting structure, i.e., semesters nested within instructors given by Model 2, we estimate δ^2, ϵ^2, and σ^2 utilizing Equation (3.12). Note that under the structure of Model 2, $\hat{\epsilon}^2$ is calculated rather than $\hat{\tau}^2$ which is a linear combination of $\hat{\delta}^2$ and all s_{ik}^2's where s_{ik}^2 (given by Equation (3.7)) represents the variance of observations within teaching method i and instructor k. A graphic of the groupings used to calculate the s_{ik}^2's for our illustration is given in Figure 3.3 indicating that $\hat{\epsilon}^2$ is calculated by taking the sample variance of all students within the same class color, then combining these variances over the two different color groups.

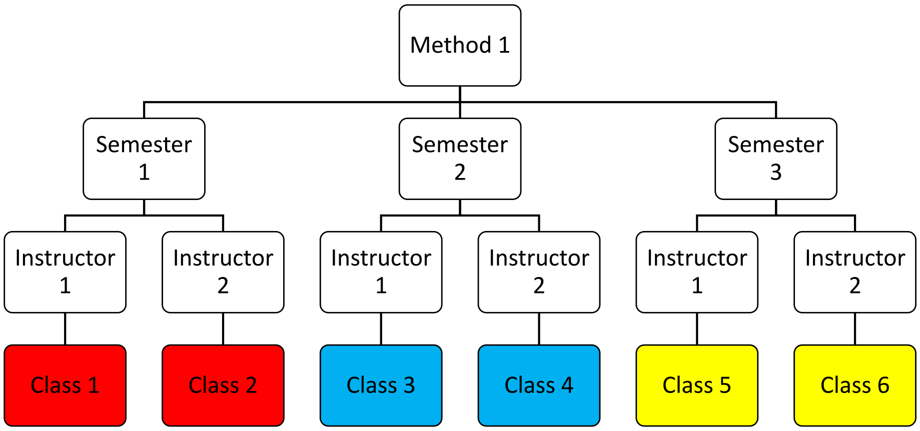


Figure 3.2 Illustration of calculating $\hat{\tau}^2$ when Model 1 is correct structure.

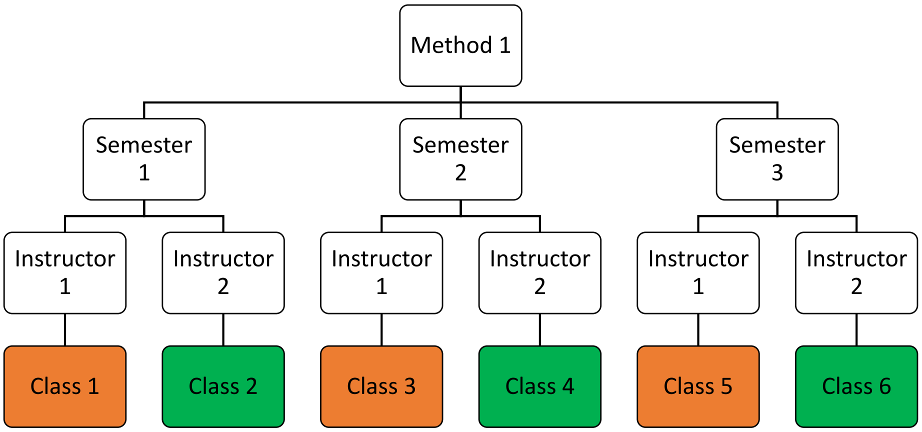


Figure 3.3 Illustration of calculating $\hat{\epsilon}^2$ when Model 1 is correct structure.

By comparing Figures 3.2 and 3.3, we can see that assuming the incorrect nesting structure (and thus estimating $\hat{\epsilon}^2$ instead of $\hat{\tau}^2$) leads to incorrect groupings of less homogeneous students. Due to this incorrect grouping of students, the variance estimate of ϵ^2 will be large, leading to a smaller estimate of σ^2, because of the sequential nature of the moment-based estimators. Thus, we expect to more frequently obtain a negative variance estimate when we assume the incorrect model structure than when we assume the correct model structure. Although in this section we have contrasted $\hat{\epsilon}^2$ and $\hat{\tau}^2$, the same phenomenon occurs with $\hat{\psi}^2$ and $\hat{\sigma}^2$. In fact, we anticipate that negative estimates of these variances will be even more common under the incorrect model structure because they are computed later in the estimation sequence of (3.12) than are $\hat{\epsilon}^2$ and $\hat{\tau}^2$.

3.3.2 Negative Moment-Based Estimates

To investigate how frequently moment-based estimates of variance components might return negative values, we conducted a simulation study in which a large number of datasets were generated and the proportion of negative moment-based variance estimators was recorded. Specifically, we simulated datasets following the structure of Model 1 (given by Equation (3.1)) which assumes instructors are nested within semesters. For simplicity, the data were simulated so that each class contained 30 student responses. Further, in all of the simulations, each teaching method had the same number of semesters as well as the same number of instructors per semester. While we maintained this balanced structure for all simulated datasets, we did vary the number of semesters and instructors within a semester to investigate the impact of different sample sizes. We considered four different sample size scenarios: three semesters per teaching method and three instructors per semester; three semesters per teaching method and ten instructors per semester; ten semesters per teaching method and three instructors per semester; ten semesters per teaching method and ten instructors per semester.

In addition to varying the number of semesters and instructors, we also varied the magnitude of the variances used to simulate the data. The variance values were chosen such that $\text{Var}(Y_{ijkl}) = \psi^2 + \tau^2 + \delta^2$ was approximately 325. (The value of 325 is based on summary statistics of exam scores for multiple introductory statistics courses at Iowa State University.) We explored three different sets of variances by considering values of ψ^2, τ^2, and δ^2 such that the differences in magnitudes were either close, moderate, or large, respectively. Table 3.1 displays the three sets of values for

ψ^2, τ^2, and δ^2. While we varied the variance parameters, the values of λ_1 and λ_2, representing the teaching method, were held constant (and equal) at sixty-five for every simulation.

Table 3.1 Three sets of variance values used to simulate data.

Parameter	Small Difference in Magnitudes	Moderate Difference in Magnitudes	Large Difference in Magnitudes
δ^2	$12^2 = 144$	$12^2 = 144$	$16^2 = 256$
τ^2	$10^2 = 100$	$11^2 = 121$	$7^2 = 49$
ψ^2	$9^2 = 81$	$8^2 = 64$	$4^2 = 16$
$\text{Var}(Y_{ijkl}) = \psi^2 + \tau^2 + \delta^2$	325	329	321

The combination of all number of instructor and semester sizes and variance values resulted in twelve distinct simulation scenarios. For each scenario we simulated 5,000 datasets and for each dataset we calculated the moment-based variance estimators, $\hat{\delta}^2$, $\hat{\tau}^2$, $\hat{\epsilon}^2$, $\hat{\psi}^2$, and $\hat{\sigma}^2$, based on Equation (3.12). We then recorded the proportion of times (out of 5,000) each moment-based estimator returned a negative estimate. The results for all twelve scenarios can be seen in Tables 3.2 through 3.4. Note that the margin of error for each calculated proportions is at most 0.014, with 95% confidence.

An examination of Table 3.2 through Table 3.4 reveals some similarities across all twelve simulation scenarios. First, the moment-based estimator of δ^2 was never less than zero. Also, the moment-based estimators of τ^2 and ϵ^2 were never less than zero for eleven of the twelve simulation scenarios. The values of Table 3.4, show that the one scenario in which $\hat{\tau}^2$ was negative occurred when there was a large difference in the variance magnitudes used to simulate the data, combined with small numbers of semesters and instructors (three each). Even then, the proportion of negative values for $\hat{\tau}^2$ was only 0.001.

Table 3.2 Proportion of times (out of 5,000 simulated datasets) that negative moment-based estimators were obtained for data simulated from Model 1, using variance parameters that have a small difference in magnitude.

	Both Models	Model 1		Model 2	
		(Correct)		(Incorrect)	
	$\hat{\delta}^2$	$\hat{\tau}^2$	$\hat{\psi}^2$	$\hat{\epsilon}^2$	$\hat{\sigma}^2$
3 Semesters, 3 Instructors	0	0	0.192	0	0.854
3 Semesters, 10 Instructors	0	0	0.037	0	0.897
10 Semesters, 3 Instructors	0	0	0.007	0	0.912
10 Semesters, 10 Instructors	0	0	0.000	0	0.968

Table 3.3 Proportion of times (out of 5,000 simulated datasets) that negative moment-based estimators were obtained for data simulated from Model 1, using variance parameters that have a moderate difference in magnitude.

	Both Models	Model 1		Model 2	
		(Correct)		(Incorrect)	
	$\hat{\delta}^2$	$\hat{\tau}^2$	$\hat{\psi}^2$	$\hat{\epsilon}^2$	$\hat{\sigma}^2$
3 Semesters, 3 Instructors	0	0	0.266	0	0.802
3 Semesters, 10 Instructors	0	0	0.058	0	0.854
10 Semesters, 3 Instructors	0	0	0.025	0	0.870
10 Semesters, 10 Instructors	0	0	0.000	0	0.927

Table 3.4 Proportion of times (out of 5,000 simulated datasets) that negative moment-based estimators were obtained for data simulated from Model 1, using variance parameters that have a large difference in magnitude.

	Both Models	Model 1		Model 2	
		(Correct)		(Incorrect)	
	$\hat{\delta}^2$	$\hat{\tau}^2$	$\hat{\psi}^2$	$\hat{\epsilon}^2$	$\hat{\sigma}^2$
3 Semesters, 3 Instructors	0	0.001	0.369	0	0.762
3 Semesters, 10 Instructors	0	0.000	0.135	0	0.774
10 Semesters, 3 Instructors	0	0.000	0.103	0	0.812
10 Semesters, 10 Instructors	0	0.000	0.001	0	0.835

Examining each of Tables 3.2 through 3.4 individually shows that as the number of semesters and instructors per semester increases, the proportion of times $\hat{\psi}^2$ is negative decreases, while the proportion of times $\hat{\sigma}^2$ is negative increases. This is true for all variance magnitudes used, indicating that as the amount of data available increases, the chances of obtaining a negative value from moment-based estimation using the correct nesting structure decrease, while those chances using the incorrect nesting structure increase. Focusing on the results across the three tables for the situations with the same sample sizes, the larger the difference in the magnitudes of variances, the more likely are negative values of $\hat{\psi}^2$ and the less likely are negative values of $\hat{\sigma}^2$. This demonstrates that, for the cases examined here, for a given sample size, negative estimates are less likely when variances are similar over levels of the hierarchy than when those differences are greater.

The results presented in Tables 3.2 through 3.4 are for data sets simulated from Model 1. Results for simulations in which data sets were generated using Model 2 are parallel, that is, the numbers of Tables 3.2 through 3.4 are similar in value, but reversed between $\hat{\tau}^2$ and $\hat{\epsilon}^2$ and also reversed between $\hat{\psi}^2$ and $\hat{\sigma}^2$.

3.3.3 A Diagnostic for Nesting Structure

The simulation study of the previous subsection suggests a diagnostic based on moment estimators of variance components. While the values of Tables 3.2 through 3.4 show marginal frequencies with which various moment-based variance estimates were negative (rather than cross-classified), those frequencies point toward the use of $\hat{\sigma}^2$, $\hat{\epsilon}^2$, $\hat{\psi}^2$, and $\hat{\tau}^2$ as the basis for determining an appropriate nesting structure. We propose a diagnostic of the following form:

1. For a given data set, estimate δ^2, τ^2, ϵ^2, ψ^2 and σ^2 using (3.12).
2. If $\hat{\sigma}^2$ or $\hat{\epsilon}^2$ is negative while $\hat{\psi}^2$ and $\hat{\tau}^2$ are both positive, choose the nesting structure of Model 1.
3. If $\hat{\psi}^2$ or $\hat{\tau}^2$ is negative while $\hat{\sigma}^2$ and $\hat{\epsilon}^2$ are both positive, choose the nesting structure of Model 2.
4. If $\hat{\sigma}^2$, $\hat{\epsilon}^2$, $\hat{\psi}^2$, and $\hat{\tau}^2$ are all positive, choose either structure, or consider a hierarchical model that is not nested (see Section 3.4).
5. If at least one $\hat{\sigma}^2$ or $\hat{\epsilon}^2$ is negative and at least one $\hat{\psi}^2$ or $\hat{\tau}^2$ is negative, then the diagnostic has failed. The structure of the data and the study from which they arose may require further scrutiny.

To examine the efficacy of this diagnostic tool, we again simulated 5,000 data sets from each of six scenarios, three in which the data were generated from Model 1 with either small (S), moderate (M), or large (L) differences in magnitude of variances, and three in which data were generated from Model 2 with either small (S), moderate (M), or large (L) differences in magnitude of variances. Variance values were the same as given in Table 3.1. All scenarios assumed ten semesters and ten instructors, the largest case examined in the previous subsection. For each data set, we computed moment-based variance estimators using Equation (3.12), and the diagnostic decision rule was applied. For ease of presentation we will call those four decisions “Choose Model 1”,“Choose

Model 2", "Choose Either Model", and "Failure". Results of this simulation exercise are presented in Table 3.3.3.

Table 3.5 Simulation results on performance of proposed diagnostic.

	Simulation					
	Model 1			Model 2		
Decision	S	M	L	S	M	L
Choose Model 1	0.969	0.931	0.842	0.000	0.000	0.001
Choose Model 2	0.000	0.000	<0.001	0.965	0.918	0.824
Choose Either Model	0.031	0.069	0.157	0.035	0.082	0.174
Failure	0.000	0.000	<0.001	0.000	0.000	<0.001

The values of Table 3.3.3 show that the proposed diagnostic has good ability to discriminate between both nesting structures. The diagnostic is not able to provide guidance in only a fraction of the datasets when the magnitudes of the difference in variances is large, corresponding to the scenario in which the data model variance used to simulate the data, δ^2, is slightly greater than five times the value of τ^2 (ϵ^2) and sixteen times the value of ψ^2 (σ^2) when simulating from Model 1 (Model 2). That is, the conditional variance of individual responses dominates all other sources of variability in the data generating mechanism. When this occurs the diagnostic indicates no preference among nesting structures in about 15% to 20% of the simulated data sets.

3.4 Non-nested Models

To this point, we have dealt exclusively with situations producing data that exhibit nested structures. Within our context of teaching methods, semesters, instructors, and classes, we have assumed that unique instructors appear nested within semesters, or unique semesters appear nested within instructors. Consider the model with instructors nested within semesters. This model treats instructors as non-replicated factors with a unique group for each semester. A model with this nesting structure could be used even if some instructors are repeated across semesters. However, in this context, there is an alternative if semesters and instructors are cross-classified, that is, if many instructors appear in multiple semesters, and vice versa. We call this structure balanced if every semester has the same set of instructors and each instructor teaches in each semester, and unbalanced otherwise. Gelman and Hill (2007) call models for this situation "non-nested", while many investigators use the phrase "linear mixed models". Using the notation of Gelman and Hill (2007), a simple non-nested model with instructors and semesters can be written as,

$$Y_{ijkl} = \lambda_i + \gamma_{j(i)} + \eta_{k(i)} + \epsilon_{jkl(i)}, \tag{3.13}$$

where we assume that the λ_i $(i = 1, \ldots, I)$ are fixed parameters and the $\epsilon_{jkl(i)}$ $(l = 1, \ldots, L_{jk(i)})$, the $\eta_{k(i)}$ $(k = 1, \ldots, K_i)$, and the $\gamma_{j(i)}$ $(j = 1, \ldots, J_i)$ are all independent such that

$$\begin{aligned} \epsilon_{jkl(i)} &\sim \mathrm{N}(0, \delta^2), \\ \eta_{k(i)} &\sim \mathrm{N}(0, \sigma^2), \\ \gamma_{j(i)} &\sim \mathrm{N}(0, \psi^2). \end{aligned} \tag{3.14}$$

The index notation $j(i)$, $k(i)$, and $jkl(i)$ in (3.13) and (3.14) indicates that the indices j and k may be re-used within each teaching method, indexed by i. As with the nested models presented earlier, this non-nested model can be extended in a straightforward manner to accommodate multiple classes taught by the same instructor in the same semester.

Model (3.13) leads to non-zero covariances for both responses having the same instructor but different semesters (i.e., the same level of k but different levels of j) and responses in the same semester but under different instructors (i.e., the same level of j but different levels of k). Specifically,

$$\begin{aligned} \mathrm{Var}(Y_{ijkl}) &= \delta^2 + \sigma^2 + \psi^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijkl'}) &= \sigma^2 + \psi^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ijk'l'}) &= \psi^2, \\ \mathrm{Cov}(Y_{ijkl}, Y_{ij'kl'}) &= \sigma^2. \end{aligned} \tag{3.15}$$

Notation for the variance components δ^2, σ^2, and ψ^2 was chosen so that the covariances in (3.15) may be compared directly with those given previously for the nested models in (3.3) and (3.4).

The implication is that, for data sets generated from this non-nested model, the moment-based estimates of σ^2, ϵ^2, ψ^2, and τ^2 of Section 3.3 should all be positive.

Additional simulations were conducted in which data were generated from model (3.13) with various differences in magnitude among the values of δ^2, σ^2 and ψ^2. Only a balanced scenario was used in which the numbers of both instructors and semesters was set equal to ten. The relative frequencies of negative values for $\hat{\sigma}^2$ and $\hat{\psi}^2$ as calculated from (3.12) and were both less than 0.001 for all combinations of variance values used in the simulations.

Simulations were also repeated in which the proposed diagnostic of Section 3.3.3 was used for each simulated data set. Data were generated from the non-nested (3.13) in four ways, which are detailed in Table 3.6.

Table 3.6 Four sets of variance values used to simulate data from model (3.13).

	Small Difference in Magnitudes		Large Difference in Magnitudes	
Parameter	$\sigma^2 = \psi^2$	$\sigma^2 < \psi^2$	$\sigma^2 = \psi^2$	$\sigma^2 < \psi^2$
δ^2	$12^2 = 144$	$12^2 = 144$	$15^2 = 225$	$15^2 = 225$
σ^2	$9^2 = 81$	$8^2 = 64$	$7^2 = 49$	$4^{=}16$
ψ^2	$9^2 = 81$	$10^2 = 100$	$7^2 = 49$	$9^2 = 81$
$\text{Var}(Y_{ijkl})$	306	308	323	322

For each of the four situations given by the columns of Table 3.6, 5,000 data sets were simulated from model (3.13). The results of applying our proposed diagnostic to these data sets are presented in Table 3.4, with a modification to the label attached to the decision resulting from cases in which $\hat{\sigma}^2$, $\hat{\epsilon}^2$, $\hat{\psi}^2$, and $\hat{\tau}^2$ are all positive. When we were considering only nested models, this case was labeled "Choose Either Model" (see Table 3.3.3). We now change that label to "Choose Non-nested Model".

Table 3.7 Simulation results on performance of proposed diagnostic when data are simulated from a non-nested model.

	Small Difference in Magnitudes		Large Difference in Magnitudes	
Decision	$\sigma^2 = \psi^2$	$\sigma^2 < \psi^2$	$\sigma^2 = \psi^2$	$\sigma^2 < \psi^2$
Choose Model 1	0.000	<0.001	0.000	0.103
Choose Model 2	0.000	0.000	0.000	0.000
Choose Non-nested Model	1.000	>0.999	1.000	0.897
Failure	0.000	0.000	0.000	0.000

The values of Table 3.4 demonstrate good performance of the proposed diagnostic when data actually arise from a non-nested situation. Combining these results with those of Table 3.3.3 shows that the proposed diagnostic is quite reliable for selection of a hierarchical model with structure appropriate for a given set of data. Performance of the diagnostic suffers some in cases for which the data model variance dominates all other sources of variability, regardless of whether the true situation involves a nested structure or not. Simulation with even our largest overall sample size, however, still involved only 100 unique instructor and semester combinations, and one might suspect the situation would improve for sets of data with a greater number of groups. In the section that follows, we will demonstrate the use of the developed diagnostic on a educational dataset.

3.5 Analysis of a Large National Educational Test

The Comprehensive Assessment of Outcomes in a First Statistics course (CAOS) is a test "designed to provide an instrument that would assess students' statistical reasoning after any first course in statistics" (delMas, 2017). The CAOS test was developed as part of the Assessment Resource Tools for Improving Statistical Thinking (ARTIST) project, a National Science Foundation funded initiative. As outlined in delMas et al. (2007), the ARTIST team developed and tested the reliability of the CAOS instrument across many iterations, eventually settling on a forty question multiple choice test. Starting in 2005, the ARTIST team made the CAOS test publicly available to instructors of introductory statistics courses and collected the student scores along with other information relating to student demographics, course variables (such as type of school the course was taught at), and the use of the CAOS test within the particular class in which it was being administered.

An observation in the dataset corresponds with a student, with each student having a unique code. For each student, we have the total score on the CAOS test (out of forty questions), as well as a code for the class the student was in and a code for the instructor the student had. Additionally, we know the type of institution (university, four-year, two-year, or high school) in which the student took the course that administered the test as well as the year and academic year in which the test was taken. By combining the information from year and academic year, we were able to determine which semester the CAOS test was taken. Note that we have data from students starting in the Fall semester of 2005 and ending in the Fall semester of 2014.

In this example, we are interested in comparing four-year schools to universities, thus, type of school will be the parameter of primary interest in terms of estimation and inference. For both types of schools, we have information from numerous instructors across many semesters with a varying number of classes for a given semester and instructor combination. Because we wish to apply the moment-based variance diagnostic developed in Section 3.3 to this dataset, we decided to limit our analysis to only classes with at least fifteen students. Similarly, we decided to include only semesters that had class results from at least three unique instructors and instructors that had class results from at least three different semesters. By restricting our analyses using these requirements, we will be able to obtain more accurate estimates of instructor and semester variances, while still maintaining a large sample size.

After subsetting the original dataset, we had data on 225 classes taught by twenty-four unique instructors across sixteen different semesters for students that were administered the CAOS test in a university course. For students given the test as part of a course at a four-year institution, we had 340 classes over eighteen semesters taught by fourteen different instructors. Because the diagnostic developed in Section 3.3 applies to scenarios in which we have one class for each instructor and semester combination, we randomly selected one class for every unique combination within a given school type. This resulted in 136 classes (taught by twenty-four instructors in sixteen semesters) at universities and ninety-one classes (taught by fourteen instructors in eighteen semesters). The final dataset includes information on 7,295 students (4,346 from a university and 2,949 from a four-year institution).

The goal is to use a model-based approach for analysis in order to determine if there is a difference in CAOS test scores for university students compared to students at four-year institutions. Due to the inherent nesting of students within classes (instructor and semester combinations) and classes within school-type, we wish to utilize a hierarchical model to account for this structure in the data. However, in order to fit a hierarchical model to the data, the specific nesting structure needs to be chosen. Based on the data collection, it is not clear whether a model with semesters nested within instructors nested within school-type is appropriate or if a model with instructors nested within semesters nested within school-type should be used. Thus, before a hierarchical model is fit, we will use moment-based variance estimator diagnostic to inform our structure choice.

3.5.1 Moment-Based Variance Estimator Diagnostic

The results from using (3.12) to obtain the moment-based variance estimators for the CAOS data are given in Table 3.8. Using the procedure outlined in Section 3.3.3, the chosen modeling structure is one in which instructors are nested within semesters, as the moment-based estimator for the variance of class means (ϵ^2) is -30.07, assuming semesters nested within instructors.

Table 3.8 Moment-based variance estimates for CAOS data.

Both Models	Model 1		Model 2	
	(Instructors in Semesters)		(Semesters in Instructors)	
$\hat{\delta}^2$	$\hat{\tau}^2$	$\hat{\psi}^2$	$\hat{\epsilon}^2$	$\hat{\sigma}^2$
153.67	47.16	3.45	-30.07	83.77

3.5.2 Hierarchical Model Formulation

Based on the results from the moment-based variance estimator diagnostic, a model with instructors nested within semesters was fit to the CAOS dataset. Specifically, define Y_{ijkl} as the CAOS exam score (as a percent) for student l ($l = 1, 2, \ldots, L_{ijk}$) taught by instructor k ($k = 1, \ldots, K_i$), in semester j ($j = 1, \ldots, J_i$), at school-type i ($i = 1$ indicates a university and $i = 2$ indicates a four-year school). Equation (3.16) gives the data model for the response and the mixing distributions are shown in (3.17). This model treats student CAOS scores as conditionally independent (given instructor, semester, and school type) Normal random variables with a class dependent mean of θ_{ijk} and a constant variance of δ^2. The class averages, θ_{ijk}, are modeled as conditionally independent (given semester and school type) Normal random variables with a semester dependent mean of μ_{ij} and a constant variance of τ^2. Finally, the semester averages, μ_{ij}, are considered conditionally independent (given the school type) Normal random variables with an expected value dependent on a fixed effect for school type, λ_i, and a constant variance, ψ^2.

$$Y_{ijkl}|\theta_{ijk} \sim \mathrm{N}(\theta_{ijk}, \delta^2). \tag{3.16}$$

$$\begin{aligned} \theta_{ijk}|\mu_{ij} &\sim \mathrm{N}(\mu_{ij}, \tau^2), \\ \mu_{ij}|\lambda_i &\sim \mathrm{N}(\lambda_i, \psi^2). \end{aligned} \tag{3.17}$$

To perform a Bayesian analysis using Markov Chain Monte Carlo methods, prior distributions were put on λ_1, λ_2, δ, τ, and ψ. The conjugate prior distributions are given in Equation (3.18), with the parameter values chosen to make the priors diffuse.

$$\begin{aligned} \lambda_1 &\sim \mathrm{N}(0, 20^2), \\ \lambda_2 &\sim \mathrm{N}(0, 20^2), \\ \delta &\sim \mathrm{Unif}(0, 20), \\ \tau &\sim \mathrm{Unif}(0, 20), \\ \psi &\sim \mathrm{Unif}(0, 20). \end{aligned} \tag{3.18}$$

3.5.3 Hierarchical Model Parameter Estimates and Inference

The package rjags in R was used to estimate the hierarchical model detailed in Section 3.5.2 using a Gibbs Sampler (Plummer, 2016). Three sets of starting values were initiated. For each set of starting values a burn-in of 5,000 observations was discarded and 20,000 observations were collected from the posterior distribution of each parameter. Mixing of the three chains was ensured by examining trace plots for individual parameters as well as obtaining the scale reduction factor (Gelman and Rubin, 1992). After determining there were no concerns with regards to the model fit, numerical summary statistics of the parameter estimates from the posterior distributions were obtained.

Table 3.9 shows five quantiles for the posterior distributions of λ_1, λ_2, and the difference $\lambda_1 - \lambda_2$, along with the three variance parameters (δ^2, τ^2 and ψ^2). Based on the Bayesian 95% credible interval for $\lambda_1 - \lambda_2$, there is evidence of a difference in CAOS scores for students at a university compared to students at a four-year school as zero is not contained in this interval. Specifically, there is a 95% chance that the mean of semester average CAOS scores for university students is between 0.385 to 5.240 percentage points larger than the mean of semester average CAOS scores for students at four-year institutions. Figure 3.4 depicts a histogram of 20,000 observations drawn from the posterior distribution of $\lambda_1 - \lambda_2$ (after a burn-in of 5,000 observations) and further emphasizes the evidence of a significant difference between the effect of school type as nearly all values are greater than zero.

Focusing on the estimates for the variance components, we see that the estimate of δ^2, which represents the variability of students within a given class, has the largest magnitude with a 95% credible interval of $(130.772, 139.571)$. The size of the δ^2 estimate is plausible, as educators identify the variance between students within a class being the largest source of variability when modeling student responses (Chance et al., 2016). While the estimate for τ^2 is smaller than that for δ^2 (the median of the posterior distribution for τ^2 is just less than half the median of the posterior distribution for δ^2), the posterior distribution of τ^2 is much more variable. With 95% probability, the variance of the class averages (given semester and school type) is between 55.405 and 82.892.

Finally, we note that the estimate of ψ^2 (a measure of the variability of semester averages within a school type) is quite small relative to the other variance values.

Table 3.9 Posterior quantiles for parameter estimates based on 20,000 observations after a burn-in of 5,000 for a hierarchical model fit to the CAOS dataset.

Parameter	2.5%	25%	50%	75%	97.5%
λ_1	55.714	56.867	57.474	58.051	59.091
λ_2	52.938	54.049	54.717	55.328	56.508
$\lambda_1 - \lambda_2$	0.385	1.902	2.706	3.576	5.240
δ^2	130.772	133.585	135.095	136.635	139.571
τ^2	55.405	62.911	67.348	72.153	82.892
ψ^2	0.002	0.097	0.452	1.319	5.366

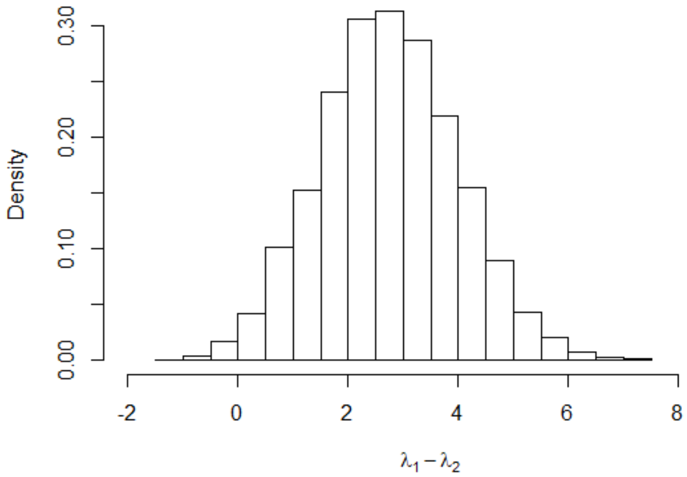


Figure 3.4 Histogram of 20,000 draws from the posterior distribution of the difference in the effects of school-type after a burn-in of 5,000.

3.5.4 CAOS Data Analysis Conclusions

In this illustration, we successfully demonstrated the use of the moment-based variance estimator diagnostic for determining the most appropriate nesting structure for modeling student responses from a large national educational test. The diagnostic led to the data-driven conclusion that a model structure in which instructors were nested within semesters (which, in-turn, were nested within school type) was the most reasonable. Based on the diagnostic results, a Bayesian hierarchical model was fit to the CAOS test scores and an examination of the posterior distributions of the parameters indicated a significant difference in the grouping variable at the top level of the hierarchy, school-type.

While there was evidence of a difference between university and four-year schools, it is important to note that, due to some restrictions of the developed diagnostic, not all available data were utilized. Further, the data analyzed were based on instructors participating voluntarily, so it would be inappropriate to extend our conclusions to all introductory statistics courses at these school-types without a more in depth look at the representativeness of the sample. Despite the cautions regarding the scope of the conclusions for this illustration, the data-driven diagnostic utilizing moment-based variance estimators proved to be a valuable tool in determining the modeling structure to use.

3.6 Discussion and Concluding Remarks

In Chapter 3 we have proposed a completely data-driven diagnostic useful in determining what type of hierarchical model might be most appropriate to reflect the structure in a set of educational data. We initially focused on models with nested structures, later extending these to include certain non-nested factors. The diagnostic appears to be reasonably accurate in identifying the correct model structure based on the results of a simulation study.

We demonstrated the usefulness of our diagnostic by applying it to a large data set that involves a well-known assessment tool for measuring educational outcomes in introductory statistics courses. In this application, the diagnostic gives a clear indication that a particular nested structure is appropriate to reflect the data, with instructors nested within semesters. Fitting this model results in estimates that indicate variability among students dominates variability due to either semesters or instructors within semesters. There appears to be a small but meaningful difference between four-year colleges and universities, which constitute the upper level effects of primary interest in this problem.

Questions remain regarding the effects that may result from fitting models that do not contain structures that are in concert with patterns in the observed data. For example, what is the effect of fitting a model with a nested structure to data that arise from a generating mechanism that would be better represented as cross-classified factors? What is the effect of the reverse situation? These questions deserve further investigation.

The hierarchical models we considered represent relatively simple structures. Extensions of the concepts introduced in this chapter should be possible for more complicated settings, such as problems that involve more levels in a hierarchy, situations in which given instructors are responsible for multiple classes in the same semester, or in which covariates are included at various levels of the hierarchy. Another issue that deserves additional investigation concerns the potential effects of unbalanced data, a phenomenon that may increase in importance as models increase in complexity.

Hierarchical models present a viable alternative for the analysis of data from education studies. These models are well suited for studies that involve a reasonably large number of individual groups of students, such as combinations of institutions, semesters, instructors, and sections of a class. At some point in the assessment of a pedagogical method, teaching technique, or organization of a course, large-scale assessments are necessary. Without such assessments, it will be impossible to draw conclusions about the effects of wide-spread adoption of a proposed approach, such as flipped classrooms or on-line instruction.

BIBLIOGRAPHY

Betihavas, V., Bridgman, H., Kornhaber, R., and Cross, M. (2016). The evidence for flipping out: a systematic review of the flipped classroom in nursing education. *Nurse Education Today*, 38:15–21.

Chance, B., Wong, J., and Tintle, N. (2016). Student performance in curricula centered on simulation-based inference: A preliminary report. *Journal of Statistics Education*, 24(3):114–126.

Clair, K. S. and Chihara, L. (2012). Team-based learning in a statistical literacy class. *Journal of Statistics Education*, 20(1):1–20.

delMas, R. (2006 (last accessed October 29, 2017)). *The CAOS test*. https://apps3.cehd.umn.edu/artisti/caos.html.

delMas, R., Garfield, J., Ooms, A., and Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2):28–58.

Geis, G. L. (1984). Comparing instructional methods: Some basic research problems. *Canadian Journal of Higher Education*, 14(2):91–98.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.

Hainey, T., Connolly, T. M., Boyle, E. A., Wilson, A., and Razak, A. (2016). A systematic literature review of games-based learning empirical evidence in primary education. *Computers & Education*, 102:202–223.

Horton, P. B., McConney, A. A., Woods, A. L., Barry, K., Krout, H. L., and Doyle, B. K. (1993). A content analysis of research published in the journal of research in science teaching from 1985 through 1989. *Journal of Research in Science Teaching*, 30(8):857–869.

Maurer, K. and Lock, D. (2016). Comparison of learning outcomes for simulation-based and traditional inference curricula in a designed educational experiment. *Technology Innovations in Statistics Education*, 9(1).

McGowan, H. M. (2011). Planning a comparative experiment in educational settings. *journal of Statistics Education*, 19(2):1–19.

McGowan, H. M. and Gunderson, B. K. (2010). A randomized experiment exploring how certain features of clicker use effect undergraduate students' engagement and learning in statistics. *Technology Innovations in Statistics Education*, 4(1).

Niehaus, E., Campbell, C. M., and Inkelas, K. K. (2014). Hlm behind the curtain: Unveiling decisions behind the use and interpretation of hlm in higher education research. *Research in Higher Education*, 55(1):101–122.

Peck, R., Olsen, C., and Devore, J. L. (2012). *Introduction to Statistics and Data Analysis*. Brooks/Cole, Belmont, CA, fourth edition.

Perrett, J. J. (2012). A case study on teaching the topic "experimental unit" and how it is presented in advanced placement statistics textbooks. *Journal of Statistics Education*, 20(2).

Plummer, M. (2016). *rjags: Bayesian Graphical Models using MCMC*. R package version 4-6.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shaver, J. P. and Norton, R. S. (1980). Randomness and replication in ten years of the american educational research journal. *Educational Researcher*, 9(1):9–15.

Tintle, N., VanderStoep, J., Holmes, V.-L., Quisenberry, B., and Swanson, T. (2011). Development and assessment of a preliminary randomization-based introductory statistics curriculum. *Journal of Statistics Education*, 19(1).

Walczak, J., Kaleta, A., Gabryś, E., Kloc, K., Thangaratinam, S., Barnfield, G., Weinbrenner, S., Meyerrose, B., Arvanitis, T. N., Horvath, A. R., et al. (2010). How are" teaching the teachers" courses in evidence based medicine evaluated? a systematic review. *BMC medical education*, 10(1):64.

Waltz, C. F., Jenkins, L. S., and Han, N. (2014). The use and effectiveness of active learning methods in nursing and health professions education: A literature review. *Nursing Education Perspectives*, 35(6):392–400.

Winquist, J. R. and Carlson, K. A. (2014). Flipped statistics class results: Better performance than lecture over one year later. *Journal of Statistics Education*, 22(3):1–10.

APPENDIX A. PROCEDURE FOR SIMULATING FROM THE POSTERIOR PREDICTIVE DISTRIBUTION FOR CHAPTER 2 DATA EXAMPLE

As referenced in Section 2.4.4, the following steps outline the procedure followed to simulate from the posterior predictive distribution of the hierarchical model found in equation (2.11) that was fit using mixing distributions and priors given in (2.9) and (2.12), respectively:

1. With the values of $\widehat{\lambda}_1$, $\widehat{\lambda}_2$, $\widehat{\psi}$, $\widehat{\alpha}$, and $\widehat{\gamma}$ simulated from the posterior distribution using the Gibbs Sampler, generate

   (a) the 17 values of σ^2_{ijk} (for each class) from an $\mathrm{IG}(\alpha, \gamma)$ distribution,

   (b) one value of μ_{11} and one value of μ_{12} from a $\mathrm{N}(\lambda_1, \psi^2)$ distribution, and

   (c) one value of μ_{21} and one value of μ_{22} from a $\mathrm{N}(\lambda_2, \psi^2)$ distribution.

2. Using the values of μ_{ij} simulated in step 1 and the value of $\widehat{\tau}$ from the posterior distribution, generate

   (a) $\beta_{0(11k)}$ from a $\mathrm{N}(\mu_{11}, \tau^2)$ for $k = 1, 2, 3, 4$,

   (b) $\beta_{0(12k)}$ from a $\mathrm{N}(\mu_{12}, \tau^2)$ for $k = 1, 2, 3, 4$,

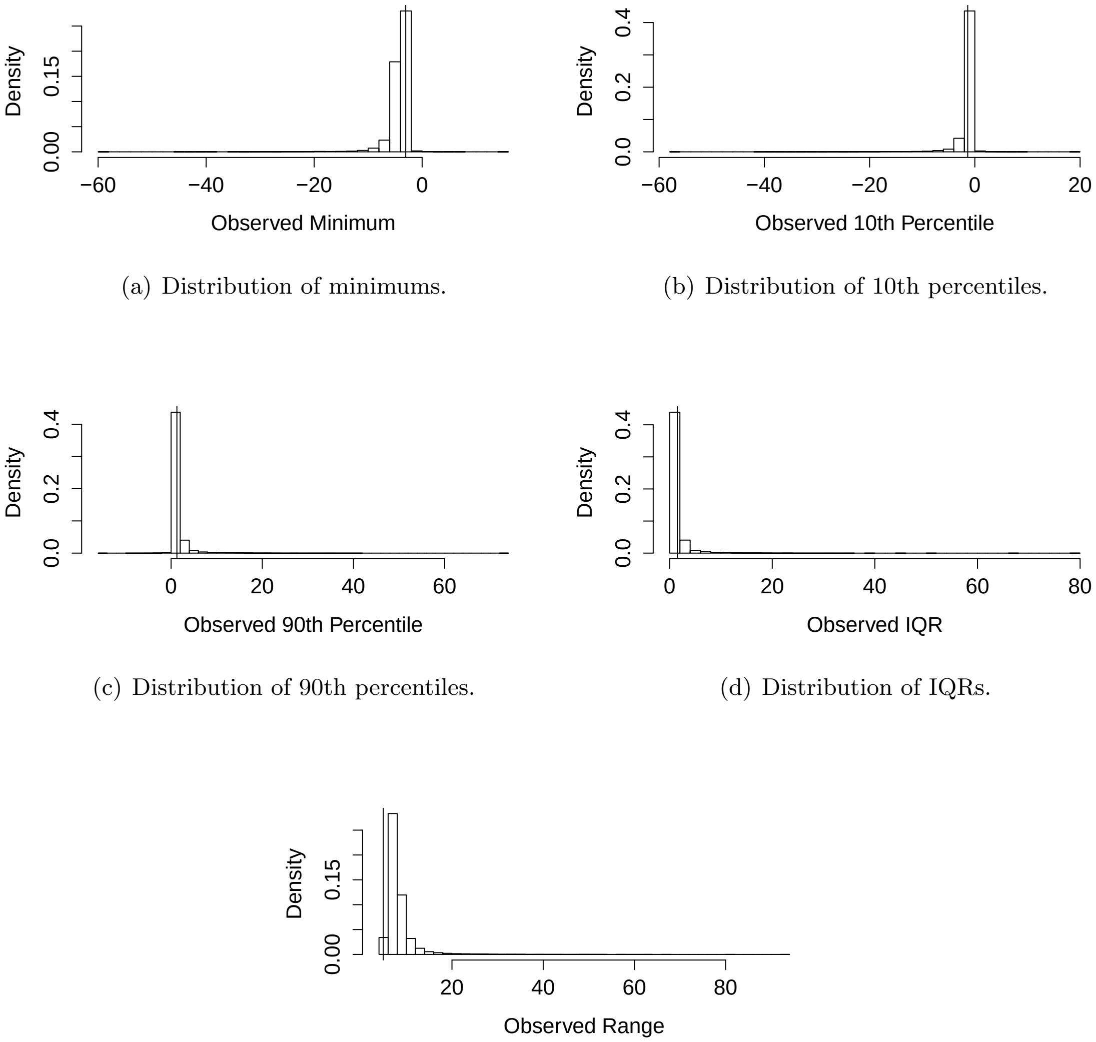
   (c) $\beta_{0(21k)}$ from a $\mathrm{N}(\mu_{21}, \tau^2)$ for $k = 1, 2, 3, 4$, and

   (d) $\beta_{0(22k)}$ from a $\mathrm{N}(\mu_{22}, \tau^2)$ for $k = 1, 2, 3, 4, 5$.

3. Using $\widehat{\beta}_1$ from the posterior distribution and the estimates from steps 1 and 2, simulate Y^*_{ijkl} from a $\mathrm{N}(\beta_{0(ijk)} + \beta_1 x_{ijkl}, \sigma^2_{ijk})$ for $i = 1, 2$, $j = 1, 2$, $k = 1, 2, 3, 4, 5$, and $l = 1, 2, \ldots, L_{ijk}$. L_{ijk} represents the number of students in term i, lecture j, and lab k. The values of Y^*_{ijkl} represent one dataset generated from the posterior predictive distribution.

APPENDIX B. ADDITIONAL PLOTS FOR CHAPTER 2 DATA EXAMPLE

Full histograms of five statistics calculated from the posterior predictive distributions simulated using the hierarchical model fit to the data referenced in Section 2.4.4.



(a) Distribution of minimums.

(b) Distribution of 10th percentiles.

(c) Distribution of 90th percentiles.

(d) Distribution of IQRs.

(e) Distribution of ranges.

APPENDIX C. PROCEDURE FOR SIMULATION BASED MODEL ASSESSMENT USING KOLMOGOROV-SMIRNOV STATISTICS FOR CHAPTER 2 DATA EXAMPLE

Below we outline the procedure for calculating Kolmogorov-Smirnov statistics based on the empircal distributions of $\widehat{\beta}_{0(ijk)}$ and $\widehat{\beta}_{1(ijk)}$ as referenced in Section 2.4.4:

1. Using the observed data, fit a linear model to each lab group individually to obtain estimates $\widehat{\beta}_{0(ijk)}$ and $\widehat{\beta}_{1(ijk)}$, $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, 3, 4, 5$.

2. Run a Gibbs sampler on the observed data to obtain M draws (used $M = 100$) from the posterior distribution of the parameters. Let $\boldsymbol{\theta_m}$ represent one such draw from the posterior distribution of the parameters. For $m = 1, 2, \ldots, M$ complete the following:

    (a) Using $\boldsymbol{\theta_m}$, simulate one dataset from the posterior predictive distribution (as described in Appendix A), call this dataset $\{Y^*_{ijkl} : i = 1, 2, j = 1, 2, k = 1, 2, 3, 4, 5, \text{ and } l = 1, 2, \ldots L_{ijk}\}$.

    (b) Fit a linear model to each lab group using the dataset generated from the posterior predictive distribution to obtain estimates $\widehat{\beta}^*_{0(ijk)}$ and $\widehat{\beta}^*_{1(ijk)}$, $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, 3, 4, 5$.

    (c) Let $G_n(\beta_0)$ be the empirical distribution function of $\widehat{\beta}_0$ and let $G^*_n(\beta_0)$ be the empirical distribution function of $\widehat{\beta}^*_0$. Calculate:

        i. $D^+_m = \sup\{G_n(\beta_0) - G^*_n(\beta_0)\}$,

        ii. $D^-_m = \sup\{G^*_n(\beta_0) - G_n(\beta_0)\}$,

        iii. $D_m = \max\{D^+_m, D^-_m\}$.

(d) Let $H_n(\beta_1)$ be the empirical distribution function of $\widehat{\beta}_1$ and let $H_n^*(\beta_1)$ be the empirical distribution function of $\widehat{\beta}_1^*$. Calculate:

  i. $C_m^+ = \sup\{H_n(\beta_1) - H_n^*(\beta_1)\}$,

  ii. $C_m^- = \sup\{H_n^*(\beta_1) - H_n^*(\beta_1)\}$,

  iii. $C_m = \max\{C_m^+, C_m^-\}$.

(e) Run a Gibbs sampler on the simulated posterior predictive data set, $\{Y_{ijkl}^* : i = 1, 2, j = 1, 2, k = 1, 2, 3, 4, 5, \text{ and } l = 1, 2, \ldots L_{ijk}\}$ to obtain P draws (used $P = 100$) from the posterior predictive distribution of the parameters. Let $\boldsymbol{\theta_p^*}$ represent one such draw from the posterior predictive distribution of the parameters. For $p = 1, 2, \ldots, P$ complete the following:

  i. Using $\boldsymbol{\theta_p^*}$, simulate one dataset from the posterior predictive distribution, call this dataset $\{Y_{ijkl}^{**} : i = 1, 2, j = 1, 2, k = 1, 2, 3, 4, 5, \text{ and } l = 1, 2, \ldots L_{ijk}\}$.

  ii. Fit a linear model to each lab group using the dataset generated from the posterior predictive distribution to obtain estimates $\widehat{\beta}_{0(ijk)}^{**}$ and $\widehat{\beta}_{1(ijk)}^{**}$, $i = 1, 2$, $j = 1, 2$, and $k = 1, 2, 3, 4, 5$.

  iii. Let $G_n^*(\beta_0)$ be the empirical distribution function of $\widehat{\beta}_0^*$ and let $G_n^{**}(\beta_0)$ be the empirical distribution function of $\widehat{\beta}_0^{**}$. Calculate:

    A. $D_{m,p}^+ = \sup\{G_n^*(\beta_0) - G_n^{**}(\beta_0)\}$,

    B. $D_{m,p}^- = \sup\{G_n^{**}(\beta_0) - G_n^*(\beta_0)\}$,

    C. $D_{m,p} = \max\{D_{m,p}^+, D_{m,p}^-\}$.

  iv. Let $H_n^*(\beta_1)$ be the empirical distribution function of $\widehat{\beta}_1^*$ and let $H_n^{**}(\beta_1)$ be the empirical distribution function of $\widehat{\beta}_1^{**}$. Calculate:

    A. $C_{m,p}^+ = \sup\{H_n^*(\beta_1) - H_n^{**}(\beta_1)\}$,

    B. $C_{m,p}^- = \sup\{H_n^{**}(\beta_1) - H_n^*(\beta_1)\}$,

    C. $C_{m,p} = \max\{C_{m,p}^+, C_{m,p}^-\}$.

3. After completing the above steps, calculate:

    (a) $D = \frac{1}{M}\sum_{m=1}^{M} D_m$,

    (b) $C = \frac{1}{M}\sum_{m=1}^{M} C_m$,

    (c) $D_m^* = \frac{1}{P}\sum_{p=1}^{P} D_{m,p}$ for $m = 1, 2, \ldots M$,

    (d) $C_m^* = \frac{1}{P}\sum_{p=1}^{P} C_{m,p}$ for $m = 1, 2, \ldots M$.

4. The values of D_m^* create a reference distribution for the statistic, D. Similarly, the values of C_m^* create a reference distribution for the statistic, C. These reference distributions can be used to check that the model is doing an adequate job generating data similar to the original dataset with respect to the distribution of slopes and the distribution of intercepts for simple linear regresion models fit to individual classes.

APPENDIX D. MOMENT-BASED VARIANCE ESTIMATORS FOR GENERIC DATASET

Let Y_{ijkl} is the response of interest for student l ($l = 1, 2, \ldots, L_{ijk}$) with instructor k ($k = 1, \ldots, K_i$) in semester j ($j = 1, \ldots, J_i$) using teaching method i ($i = 1, 2$). Thus, K_1 represents the number of instructors in teaching method one and K_2 represents the number of instructors in teaching method two. Similarly, J_1 and J_2 represent the number of semesters in teaching method one and two, respectively. Define C to be the total number of distinct classes, where each class has a unique method, semester, and instructor combination. Under this generic situation, the moment based variance estimators from Equation 3.12 have the following form:

$$\widehat{\lambda}_1 = M_1,$$

$$\widehat{\lambda}_2 = M_2,$$

$$\widehat{\delta}^2 = \left(\frac{1}{C}\right) \sum_i \sum_j \sum_k \left(s^2_{ijk}\right),$$

$$\widehat{\tau}^2 = \left(\frac{1}{J_1 + J_2}\right) \sum_i \sum_j \left[\left(\frac{\left(\sum_k L_{ijk}\right)^2 - \sum_k L_{ijk}}{\left(\sum_k L_{ijk}\right)^2 - \sum_k L^2_{ijk}} \right) \left(s^2_{ij} - \widehat{\delta}^2 \right) \right],$$

$$\widehat{\epsilon}^2 = \left(\frac{1}{K_1 + K_2}\right) \sum_i \sum_k \left[\left(\frac{\left(\sum_j L_{ijk}\right)^2 - \sum_j L_{ijk}}{\left(\sum_j L_{ijk}\right)^2 - \sum_j L^2_{ijk}} \right) \left(s^2_{ik} - \widehat{\delta}^2 \right) \right],$$

$$
\begin{aligned}
\widehat{\psi}^2 = {} & \left(\frac{\sum_i \sum_j \sum_k L_{ijk}}{\left(\sum_i \sum_j \sum_k L_{ijk}\right)^2 - \sum_i \sum_j \left(\sum_k L_{ijk}\right)^2} \right) \Bigg[\left(\sum_i \sum_j \sum_k L_{ijk} - 1 \right) s^2_{ijkl} \\
& - \left(\sum_j \sum_k L_{1jk} - \frac{\left(\sum_j \sum_k L_{1jk}\right)^2}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_1^2 - \left(\sum_j \sum_k L_{2jk} - \frac{\left(\sum_j \sum_k L_{2jk}\right)^2}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_2^2 \\
& + \left(\frac{2\left(\sum_j \sum_k L_{1jk}\right)\left(\sum_j \sum_k L_{2jk}\right)}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_1 \widehat{\lambda}_2 - \left(\sum_i \sum_j \sum_k L_{ijk} - 1 \right) \widehat{\delta}^2 \\
& - \left(\sum_i \sum_j \sum_k L_{ijk} - \frac{\sum_i \sum_j \sum_k L^2_{ijk}}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\tau}^2 \Bigg],
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\sigma}^2 = {} & \left(\frac{\sum_i \sum_j \sum_k L_{ijk}}{\left(\sum_i \sum_j \sum_k L_{ijk}\right)^2 - \sum_i \sum_k \left(\sum_j L_{ijk}\right)^2} \right) \Bigg[\left(\sum_i \sum_j \sum_k L_{ijk} - 1 \right) s^2_{ijkl} \\
& - \left(\sum_j \sum_k L_{1jk} - \frac{\left(\sum_j \sum_k L_{1jk}\right)^2}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_1^2 - \left(\sum_j \sum_k L_{2jk} - \frac{\left(\sum_j \sum_k L_{2jk}\right)^2}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_2^2 \\
& + \left(\frac{2\left(\sum_j \sum_k L_{1jk}\right)\left(\sum_j \sum_k L_{2jk}\right)}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\lambda}_1 \widehat{\lambda}_2 - \left(\sum_i \sum_j \sum_k L_{ijk} - 1 \right) \widehat{\delta}^2 \\
& - \left(\sum_i \sum_j \sum_k L_{ijk} - \frac{\sum_i \sum_j \sum_k L^2_{ijk}}{\sum_i \sum_j \sum_k L_{ijk}} \right) \widehat{\epsilon}^2 \Bigg]
\end{aligned}
$$

APPENDIX E. MOMENT-BASED VARIANCE ESTIMATORS FOR COMPLETELY BALANCED DATASET

Let Y_{ijkl} be the response of interest for student l ($l = 1, 2, \ldots, L$) with instructor k ($k = 1, \ldots, K$) in semester j ($j = 1, \ldots, J$) using teaching method i ($i = 1, 2$). Here, we are assuming we have two teaching methods and within each teaching method there are J semesters for each of K instructors. A unique teaching method, instructor, and semester combination constitutes one class and within a class there are a total of L students. Thus, we have a completely balanced situation with a total of $I * J * K * L$ students across $I * J * K$ classes. Under this balanced situation, the moment based variance estimators from Equation 3.12 have the following form:

$$\widehat{\lambda}_1 = M_1,$$

$$\widehat{\lambda}_2 = M_2,$$

$$\widehat{\delta}^2 = \left(\frac{1}{IJK}\right) \sum_i \sum_j \sum_k \left(s_{ijk}^2\right),$$

$$\widehat{\tau}^2 = \left(\frac{1}{2J}\right) \sum_i \sum_j \left[\left(\frac{KL-1}{L(K-1)}\right)\left(s_{ij}^2 - \widehat{\delta}^2\right)\right],$$

$$\widehat{\epsilon}^2 = \left(\frac{1}{2K}\right) \sum_i \sum_k \left[\left(\frac{JL-1}{L(J-1)}\right)\left(s_{ik}^2 - \widehat{\delta}^2\right)\right],$$

$$\widehat{\psi}^2 = \left(\frac{1}{KL(IJ-1)}\right)\left[(IJKL-1)s_{ijkl}^2 - JKL\left(\frac{I-1}{I}\right)\widehat{\lambda}_1^2 - JKL\left(\frac{I-1}{I}\right)\widehat{\lambda}_2^2\right.$$

$$\left. + (JKL)\widehat{\lambda}_1\widehat{\lambda}_2 - (IJKL-1)\widehat{\delta}^2 - L(IJK-1)\widehat{\tau}^2\right],$$

$$\widehat{\sigma}^2 = \left(\frac{1}{JL(IK-1)}\right)\left[(IJKL-1)s_{ijkl}^2 - JKL\left(\frac{I-1}{I}\right)\widehat{\lambda}_1^2 - JKL\left(\frac{I-1}{I}\right)\widehat{\lambda}_2^2\right.$$

$$\left. + (JKL)\widehat{\lambda}_1\widehat{\lambda}_2 - (IJKL-1)\widehat{\delta}^2 - L(IJK-1)\widehat{\epsilon}^2\right],$$